Objective Scaling of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners

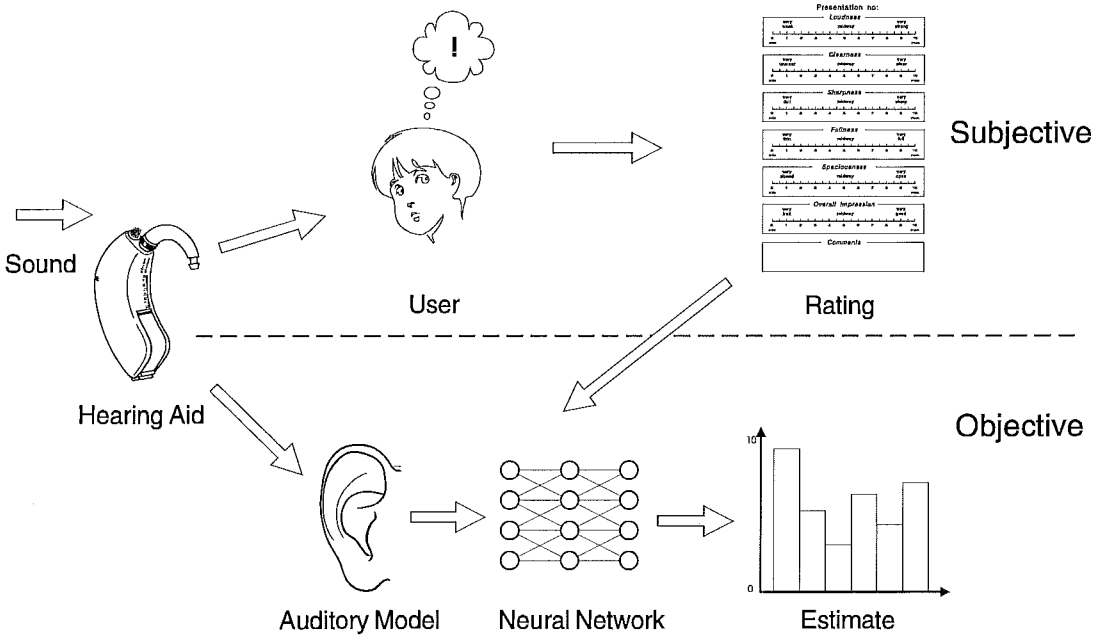
Nielsen, Lars Bramsløw

Publication date:
1993

Document Version
Publisher's PDF, also known as Version of record

ISSN 0105-3027

Objective Scaling of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners.



THE ACOUSTICS LABORATORY

TECHNICAL UNIVERSITY OF DENMARK

Report No. 54, 1993

Objective Scaling of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners.

by

Lars Bramsløw Nielsen

Oticon Research Unit "Eriksholm"

and

The Acoustics Laboratory
Technical University of Denmark

Submitted in partial fulfillment of the Danish Ph.D. degree. This work was sponsored by Oticon A/S and the Academy of the Technical Sciences (ATV), Denmark:

Project number: EF 356.

Project title: Modeling sound quality for hearing-impaired listeners.

Thesis overview.

This summary report is submitted in partial fulfillment of the Danish Ph.D. degree. The complete Ph.D. project, entitled "Modeling Sound Quality for Hearing-Impaired Listeners" is documented in four reports, of which this is the final one. Together, they form the Ph.D. thesis. In the following, the other reports are referred to as Report 1 - Report 3, and the respective Abstracts and Tables of Contents are reproduced in Appendices 13.1 - 13.6.

Report 1:

Nielsen, Lars Bramsløw (1992). **Subjective Evaluation of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners.** Internal report no. 43-8-1, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 51, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

Report 2:

Nielsen, Lars Bramsløw (1993a). **An Auditory Model with Hearing Loss.** Internal report no. 43-8-2, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 52, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

Report 3:

Nielsen, Lars Bramsløw (1993b). **A Neural Network Model for Prediction of Sound Quality.** Internal report no. 43-8-3, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 53, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

This Report:

Nielsen, Lars Bramsløw (1993c). **Objective Scaling of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners.** Internal report no. 43-8-4, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 54, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

Printed by: The Acoustics Laboratory, Technical University of Denmark
August 1993.

Abstract.

A new method for the objective estimation of sound quality for both normal-hearing and hearing-impaired listeners has been presented: OSSQAR (Objective Scaling of Sound Quality and Reproduction).

OSSQAR is based on three main parts, which have been carried out and documented separately: **1**) Subjective sound quality ratings of clean and distorted speech and music signals, by normal-hearing and hearing-impaired listeners, to provide reference data, **2**) An auditory model of the ear, including the effects of hearing loss, based on existing psychoacoustic knowledge, coupled to **3**) An artificial neural network, which was trained to predict the sound quality ratings.

OSSQAR predicts the perceived sound quality on two independent perceptual rating scales: Clearness and Sharpness. These two scales were shown to be the most relevant for assessment of sound quality, and they were interpreted the same way by both normal-hearing and hearing-impaired listeners. The scales were found not to be absolute, thus OSSQAR cannot predict the absolute sound quality, but it can be used to rank the sound quality.

Using test data from the present subjective rating experiment, the prediction error of OSSQAR was found to be only slightly larger than the random variance in the subjective ratings. Analysis of the neural network after training could not identify qualitative relations between the physical signal parameters and the perceived sound quality.

Further verification of OSSQAR with new signals and distortion types will be required to assess how general and reliable OSSQAR is, and to identify the limitations of its application.

Keywords: Sound quality, objective, hearing loss, auditory model, neural network.

Resumé.

En ny metode til objektiv prediktion af lydkvalitet for såvel normalthørende som hørehæmmede lyttere er blevet forelagt: OSSQAR (Objective Scaling of Sound Quality And Reproduction: Objektiv skalering af lydkvalitet og lydgengivelse).

OSSQAR er baseret på tre hoved-elementer, som er blevet udført og dokumenteret hver for sig: **1**) Subjektiv vurdering af lydkvaliteten af rene og forvrængede tale- og musik-signaler, udført af normalthørende og hørehæmmede lyttere, til fremskaffelse af reference data, **2**) En auditiv model af øret, der inkluderer virkningen af høretab udfra den eksisterende psykoakustiske viden, sammenkoblet med **3**) Et kunstigt neuralt netværk, som blev trænet til at prediktere den vurderede lydkvalitet.

OSSQAR predikterer den opfattede lydkvalitet på to uafhængige subjektive skalaer: Klarhed og Skarphed. Disse to skalaer viste sig at være de mest relevant for vurdering af lydkvalitet, og de blev fortolket ens af de normalthørende og de hørehæmmede lyttere. Skalaerne viste sig ikke at være absolutte, det vil sige at OSSQAR ikke kan forudsige den absolutte lydkvalitet, men kan benyttes til en rangordning af lydkvalitet.

Ved brug af test data fra nærværende lytteforsøg, blev det påvist at prediktions-fejlen for OSSQAR kun var lidt større end den statistiske usikkerhed i lydkvalitets vurderingerne. Analyse af det trænede neurale netværk, kunne ikke identificere kvalitative relationer mellem de fysiske signalparametre og den opfattede lydkvalitet.

Yderligere verifikation af OSSQAR med nye signaler og forvrængnings-typer vil være nødvendige for at bestemme generaliteten og pålideligheden af OSSQAR, og for at vurdere grænserne for dens anvendelse.

Stikord: Lydkvalitet, objektiv, høretab, auditiv model, neuralt netværk.

Preface.

This report is submitted in partial fulfillment of the Ph.D. degree. It has been written as the final report in the Ph.D. project, entitled "Modeling Sound Quality for Hearing-Impaired Listeners", which was completed in the period 1.02.1991 to 31.07.1993. The project was a joint project between Oticon A/S, The Acoustics Laboratory, Technical University of Denmark and The Speech Technology Center, Aalborg University, Denmark. It was sponsored by Oticon A/S and the Academy of the Technical Sciences (ATV), Denmark, under project number EF 356. The present report has thus been published by the two main partners: Oticon Internal Report No. 43-8-4 and The Acoustics Laboratory, Report no. 54.

The present report is the main part of the thesis, summarizing the entire project with enough details to stand alone. The three companion reports, Report 1 (Nielsen, 1992), Report 2 (Nielsen, 1993a), and Report 3 (Nielsen, 1993b) have also been written to stand alone. Therefore, there is a certain amount of overlap between the four reports, as the enthusiastic reader will notice. Throughout the text, terms from psychoacoustics, signal processing and other areas are used. The unfamiliar reader is recommended to study the three companion reports, and perhaps the background literature cited there.

The present report is organized as follows: **Section 1** provides an introduction to different types of sound quality measures, subjective and objective, and a classification of current measures is presented. **Section 2** reviews some of the main literature in the field, providing a starting point for the present work. **Section 3** outlines the present project, termed OSSQAR (Objective Scaling of Sound Quality And Reproduction). **Section 4** summarizes the first study, designed to obtain subjective ratings of sound quality for normal-hearing and hearing-impaired listeners (Report 1). **Section 5** is a summary of an auditory model with hearing loss, which was developed as a pre-processor for the objective sound quality measure (Report 2). **Section 6** describes a neural network model that was successfully trained and used to predict sound quality

based on the output from the auditory model (Report 3). **Section 7** evaluates the prediction performance of OSSQAR and discusses the utility of the measure. **Section 8** discusses the relation between the objective (physical) parameters and the subjective perception of sound quality. **Section 9** is a discussion of the outcomes and implications of the present work, and **section 10** is the conclusion of the project.

On a personal note, this has been an extremely multi-disciplinary project including elements of: auditory physiology, psychoacoustics, psychophysical procedures, neural networks in theory and practice, speech recognition, statistics, computer science, signal processing, electroacoustics etc. Such a diverse project was challenging, but from time to time also frustrating, since there was little time to venture further into any particular area. Much of the work was in the direction of the 'soft' sciences, which sometimes made it impossible to provide definite 'yes' or 'no' answers. But then, these are the conditions you have to face when you deal with the real world! I have learned that research continuously provides one with more and more insight into one's own lack of knowledge, and that a project like this in some ways is more of a mental challenge than an engineering challenge.

I want to acknowledge some of the people, who have helped me during the project. My advisors: Claus Elberling, Oticon A/S, Torben Poulsen, The Acoustics Laboratory and Paul Dalsgaard, Center for Speech Technology, Aalborg University Center. Together, they tried to keep my feet just a little bit on the ground, and asked some of those annoying questions like some researchers do. My colleagues at the Oticon research center "Eriksholm" and at the Acoustics Laboratory. All of my friends and my family who have lent their ears to some grief here and there. OLE, from Windows, who taught me everything I never wanted to know about WYGINWYW software (What You Get Is Not What You Want). And finally, I want to thank myself, without whom this work would never have been completed.

Lars Bramsløw Nielsen
Snekkersten, July 1993

Table of contents.

1. Introduction. 11
1.1 Subjective measures of sound quality. 12
1.2 Objective measures of sound quality. 13
1.3 Classification of sound quality measures. 14
2. Review of existing sound quality measures. 19
2.1 Relative sound quality measures. 19
2.2 Absolute sound quality measures. 30
3. OSSQAR: Project scope and goals. 37
4. OSSQAR: Subjective measures. 41
4.1 Purpose 41
4.2 Materials and methods. 41
4.3 Main results. 45
4.4 Conclusion: Subjective sound quality. 51
5. OSSQAR: Auditory modeling. 53
5.1 Purpose and motivation. 53
5.2 Model description. 55
5.3 Verification. 58
5.4 Conclusion: Auditory model. 59
6. OSSQAR: Neural network model. 61
6.1 Data representation. 61
6.2 Neural network implementation. 63
6.3 Test sets and performance. 64
6.4 Analysis of weights. 65
6.5 Conclusion: Neural network model. 65
7. Evaluation of OSSQAR. 67
7.1 Design choices. 67
7.2 Verification with test set data. 68
7.3 Prediction performance. 70
7.4 Discussion: OSSQAR 76

8. From the physical to the subjective domain. 81
8.1 Signal and processing effects (factorial analysis). 81
8.2 Auditory model and neural network interpretation. 84
9. Discussion. 87
10. Conclusion. 91
11. Suggestions for future work. 93
12. References. 97
13. Appendices. 101
13.1 Abstract - Report 1. 101
13.2 Table of contents - Report 1. 102
13.3 Abstract - Report 2. 104
13.4 Table of contents - Report 2. 104
13.5 Abstract - Report 3. 106
13.6 Table of contents - Report 3. 107

1 Introduction.

Sound quality remains a very subjective and poorly understood aspect of sound reproduction. At the same time, everyone has an idea of 'good' sound quality and 'poor' sound quality and they will encounter it everywhere in modern society, when listening to the radio, to the home stereo set, when using a telephone and for the hearing-impaired populations, when listening though an amplification device, i.e. a hearing aid.

The most important requirement for a hearing aid is to amplify speech sounds, in order to make them intelligible and alleviate the communication disorder of a hearing-impaired person. It is also crucial for the user, that the sound quality is natural and pleasant for speech sounds and other sound sources, such as music and environmental sounds. Due to limitations in size and power consumption combined with the requirement for high sound power output a compromise must often be made on sound quality. It is also not clear how much the hearing impairment itself affects the perception of sound quality, but the fitting of the hearing aid to an individual's hearing loss is a difficult task, where sound quality is a very important consideration. In the development phase of a hearing aid, sound quality should also be optimized, but the guidelines for this are often very unclear.

The definition of sound quality is discussed in a previous report on subjective evaluation of sound quality (Report 1), and the following definition was used throughout the present project:

> *"The **sound quality** of a hearing aid are those attributes in the auditory perception that describe the timbre and naturalness of the reproduction. Speech intelligibility is not part of sound quality. Sound quality becomes meaningful only when most sounds have been made audible without being uncomfortably loud."*

1.1 Subjective measures of sound quality.

Traditional assessments of sound quality are conducted as time-consuming, expensive laboratory experiments, which make a close control of all experimental parameters possible. This situation is often unrealistic and does not put the hearing-aid user in typical environments. The task of the user is then typically to rate the sound quality on a number of perceptual scales, e.g. Clearness, Sharpness, Pleasantness, Naturalness, Fidelity etc., or to make comparison of reproductions in pairs and select the preferred reproduction. Another problem with laboratory experiments is the limited exposure time compared to daily use (Rydén, 1993). The subject will not become more tolerant (or more annoyed), as in daily use. An alternative is to use a field-test, where the user must assess the sound quality during daily use in a variety of situations (dialog in quiet and noise, music listening, daily chores etc.) and respond by means of a questionnaire. This approach is also time-consuming and often very inaccurate and insensitive, due to potential subject bias, poor control of the acoustic environment and other factors. See Report 1 for a detailed discussion of subjective evaluation techniques.

In order to produce reliable results from subjective listening tests they have to be planned and carried out carefully and the results must be analyzed using statistical methods. If the research questions posed have not been answered, due to poor problem formulation, poor experimental design etc., the experiment must be modified and repeated. This process is acceptable for instance for an evaluation of a new product prior to market release, but it is not feasible to use in the development process every time a "knob is turned" or a circuit or an algorithm is modified. Often a hearing aid developer will then resort to his/her own (normal) ear in this process and evaluate sound quality in quick, informal sessions, where the listening situation is often artificial. This is a biased method, since the developer is aware of the signal processing taking place and may listen selectively for particular types of distortion. A personal interest in quick progress may also color the judgment of the developer.

1.2 Objective measures of sound quality.

The traditional technical (objective) measurements carried out on a hearing aid include frequency responses for a range of input levels, harmonic and intermodulation distortion and internal noise, but the perceptual effect of these is poorly understood, and only extreme cases can be flagged as being obviously imperfect. For the modern non-linear hearing aids, some of the traditional measurements, such as a swept pure tone frequency response, can be quite misleading. Recently, coherence measurements by means of two-channel FFT analysis has been introduced and discussed for objective evaluation of hearing aids (Dyrlund, 1992; Kates, 1992), which separated the linear (correlated) part of the response from the non-linear (un-correlated, like noise and distortion) part of the response. There are, however, no perceptual criteria for this coherence - what is "good" and "bad"? The coherence method separates linear and non-linear response rather than distinguishing between desired and undesired signal processing in the hearing aid. Even this distinction has no meaning to the listener, who simply judges the overall sound quality either by itself or by comparing to other hearing aids in a more or less formal manner.

Recently, most of the activity in developing objective measures, has been for evaluation of the new bit-rate reduction coders and decoders (codecs). These codecs exploit the masking properties of the human ear to eliminate masked bands of the signal and to quantize with the least number of bits required to keep quantization noise inaudible. The optimal performance for such a codec is not transparency but "imperceptibility", so the traditional signal measures fail to reveal imperfections in the codec (false negative), or they may detect degradations that are not audible (false positive). Instead, subjective and objective evaluations of the degradation of the original signal are used. For the previously mentioned reasons, objective measures are attractive: They provide repeatable results faster. Two principally different types of measures are used, both with reference to the unprocessed signal: Audibility of artifacts, i.e. a threshold measure, or rating of the relative degradation on an impairment scale (not to be confused with hearing impairment as in hearing loss). These measures are not directly

applicable to hearing aids, because no optimal reference is available. In a hearing aid, frequency shaping and sometimes dynamic range compression are desired types of signal processing, and transparency to the user is not necessarily the optimum.

Of course, the crucial and difficult question to answer for all objective quality measures is how well they represent the listener population that they are aimed for. How well do they predict sound quality and are there conditions, signals or listeners that cause them to fail? As we shall see in section 2, quite a few objective measures have been proposed on technical or psychoacoustical grounds, but much work still remains in the verification process.

It is also important to note, that no objective quality measures to date have been proposed for, or applied to hearing aids. The present report presents the first example of such a measure.

1.3 Classification of sound quality measures.

Based on the above considerations, we can define criteria for making a classification of the existing subjective and objective quality measures:

- Is the measure relative or absolute? With a relative measure, each signal condition to be measures is compared to some other condition, either a perfect reference, or all other conditions. The outcome thus depends on the other conditions in the experiment, and is not reproducible if the conditions have changed. With an absolute measure, the rating of one condition is in principle independent of the other conditions and requires no comparison - a hearing aid can for instance be rated to have a Clearness of 7 on a 0 - 10 scale. These "absolute" measures are in practice context-dependent and thus not truly absolute (Report 1).

- Is the measure a threshold or is it numerical? A threshold measure determines if some signal processing or degradation is audible or not. A numerical measure provides some metric for the degree of degradation or modification. This metric can be a ratio scale (known point, e.g. zero) or an interval scale without a known reference point. See Report 1 for a discussion of the scale types.

- Does the measure have a known optimum? A degradation measure will usually have an optimal point, i.e. no audible change. A rating scale of Overall Impression will usually also have a perceptual optimum, i.e. 10 on a 0 - 10 scale. Or the optimum is located at the center of the scale, e.g. midway on the Loudness scale (neither too Loud nor to Soft). A Sharpness scale, on the other hand, may not have an obvious optimum.

- For the objective measures it is important to consider what the subjective counterpart is. For instance, an objective measure that determines the audibility of some signal processing has a subjective counterpart in a paired comparison experiment, where the degradation is indicated by a yes/no answer, i.e. a threshold experiment.

There are many flavors of experiments and methods, but the following overview may provide some clarification. The subjective and objective sound quality measures discussed in the present report and in (Report 1) are summarized in Figures 1 and 2, according to the criteria mentioned above. The measures presented in the literature will be discussed further in section 2, while the new proposed measure, OSSQAR (Objective Scaling of Sound Quality and Reproduction) is discussed in the remainder of the present report.

SUBJECTIVE MEASURES OF SOUND QUALITY.

Type of measure Literature reference	Absolute or relative measure.	Threshold or numerical measure.	Known optimum
Paired comparison with preference judgment. Bech, 1987	Relative. All conditions are combined in pairs.	Threshold - numerical distance measure can be derived.	No - depends on the other conditions.
Paired comparison with similarity rating. Punch et al, 1980	Relative to overall mean.	Numerical distance measure.	No - depends on the other conditions.
Paired comparison with fixed reference.	Relative to known reference.	Threshold.	Yes, the transparent system.
Paired comparison with similarity rating to fixed reference. Grewin, 1993	Relative to known reference.	Numerical - ratio scale, i.e. degradation.	Yes, the transparent system.
Adjective rating. Quackenbush et al, 1988 Report 1 (Nielsen, 1992) Gabrielsson & Hagerman, 1993	Absolute within experiment. To some extent also outside exp.	Numerical - interval scale.	For certain rating scales. Subjects can also be asked to provide ideal ratings.
Informal listening tests.	Both. Comparisons are often used. Often biased.	Neither, since no formal experiment is conducted.	No, and subject is not representative of a population.

1. Table of subjective measures and methods for the evaluation of sound quality. Compare to objective measures in Figure 2. See section 2 for details.

Some of the subjective measures listed above have been included for completeness, although they have not been used for development of a corresponding objective measures.

OBJECTIVE MEASURES OF SOUND QUALITY.

Type of measure Literature reference	Absolute or relative measure.	Threshold or numerical measure.	Known optimum	Subjective counterpart.
Frequency response	Absolute.	Not related to perception.	Not for hearing aids.	None.
Noise + distortion	Absolute.	Not related to perception.	As little as possible for linear HA.	None.
Coherence Dyrlund, 1992 Kates, 1992	Absolute.	Not related to perception.	Yes, for linear hearing aids.	None.
Auditory Spectrum Distance (ASD) Karjalainen, 1985	Relative to transparent.	Numerical.	Auditory Spectrum Distance = 0.	Absolute adjective rating with fixpoints on scale.
Distortion index Levitt et al, 1987a	Relative to estimated linear system.	Threshold.	Zero distortion.	Paired comparison with original signal - threshold test.
Composite Acceptability Quackenbush et al, 1988	Relative to transparent.	Numerical	For some scales	DAM - Diagnostic Acceptability (adjective rating).
Noise-to-Masker Ratio (NMR) and Masking Flag Brandenburg, 1993	Relative to transparent. Error signal is calculated.	Threshold (Audibility flag) and margin (dB).	NMR << 0 dB (the transparent system)	Paired comparison with original signal - threshold test.
PERCEVAL Paillard, et al 1992	Relative to transparent.	Threshold.	Zero probability of detection.	Paired comparison with original signal - threshold test.
PAQM Beerends & Stemerdink, 1992	Relative to transparent. Internal error is calculated.	Numerical - can predict impairment score.	The transparent system.	Comparison rating with fixed reference.
Sharpness von Bismarck, 1974b	Absolute.	Numerical	No	Paired comparison with preference - equal, half or double Sharpness
Pleasantness. Zwicker & Fastl, 1990	Absolute.	Numerical	No.	Comparison rating with known reference.
OSSQAR This Report	Absolute, but context-dependent	Numerical	For some scales.	Adjective rating.

2. *Table of objective measures and methods for the evaluation of sound quality. Compare to subjective measure in Figure 1. See section 2 for details.*

2 Review of existing sound quality measures.

The purpose of this section is to present the current state of objective measures, and to attempt to systematize the types of measures. This is important both for subjective and objective measures, since there are meaningful correspondences between the two domains in some cases. And it is crucial to be aware of this correspondence, so that the results of objective measures are not interpreted beyond their limitations. Therefore, the objective measures in Figure 2 have been listed with their subjective counterpart (listed in Figure 1) to the extent that the original authors were specific about this. Otherwise, the present author has made a judgment of the appropriate match.

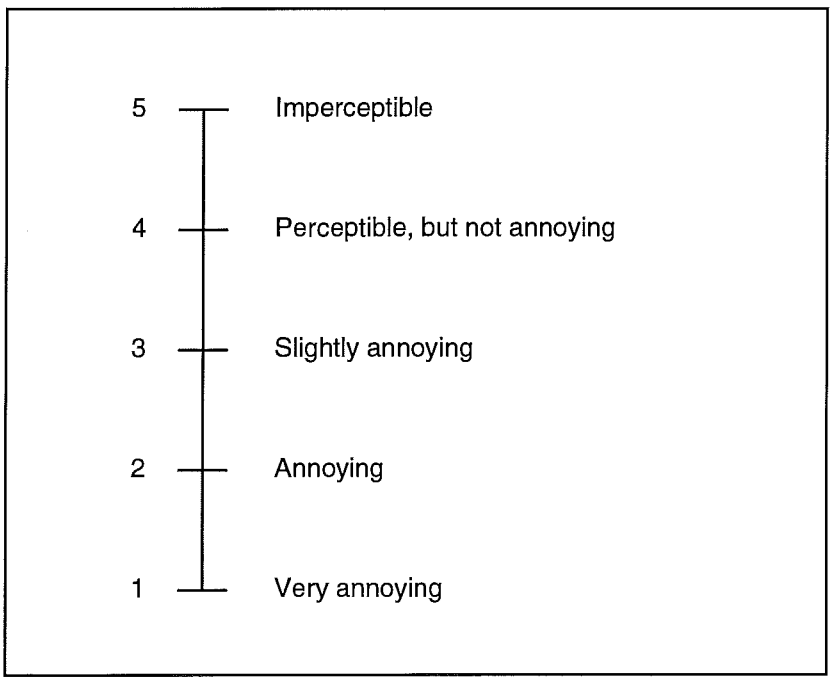
Readers mainly interested in the details concerning OSSQAR may skip this section.

2.1 Relative sound quality measures.

See figure 1 for an overview of relative subjective measures. These are always with reference to some other condition(s). In standard paired comparison experiments, one condition (i.e. one reproduction) is compared against all other conditions.

In paired comparison with preference judgment the subject must indicate the best condition with respect to some perceptual attribute - e.g. which reproduction is the Clearer of the two. Or simply indicate which condition provides the better sound quality (Overall Impression). When all pairs have been compared, a distance measure can be derived from the preference judgments (Bech, 1987), that expresses the distance to an overall average. There is no optimum, but an overall winner can typically be found based on preference statistics or the distance measure. Paired comparison with similarity rating requires the subject to rate the distance (or similarity) between the two conditions in a pair, according to a perceptual attribute (e.g. Clearness) or no attribute (e.g. "how similar are they?" - Punch et al (1980)).

Paired comparison can also be with a fixed reference. The threshold variant is paired comparison with fixed reference, where the subject must indicate "is there a change?" or "is there a degradation?". The numerical measure variant is paired comparison with similarity rating to fixed reference, where the subject compares the different signal conditions against a fixed reference, typically the original signal. A recent, often cited application for this is for rating of bit-rate reduction codecs on the CCIR "5-grade impairment" scale, that expresses degradation relative to perfect (= 5) on the scale (Grewin, 1993). This can be considered a ratio scale (defined in Report 1), because there is a well-defined reference point, namely the transparent reproduction. An example of the CCIR 5-grade impairment scale is shown in Figure 3.

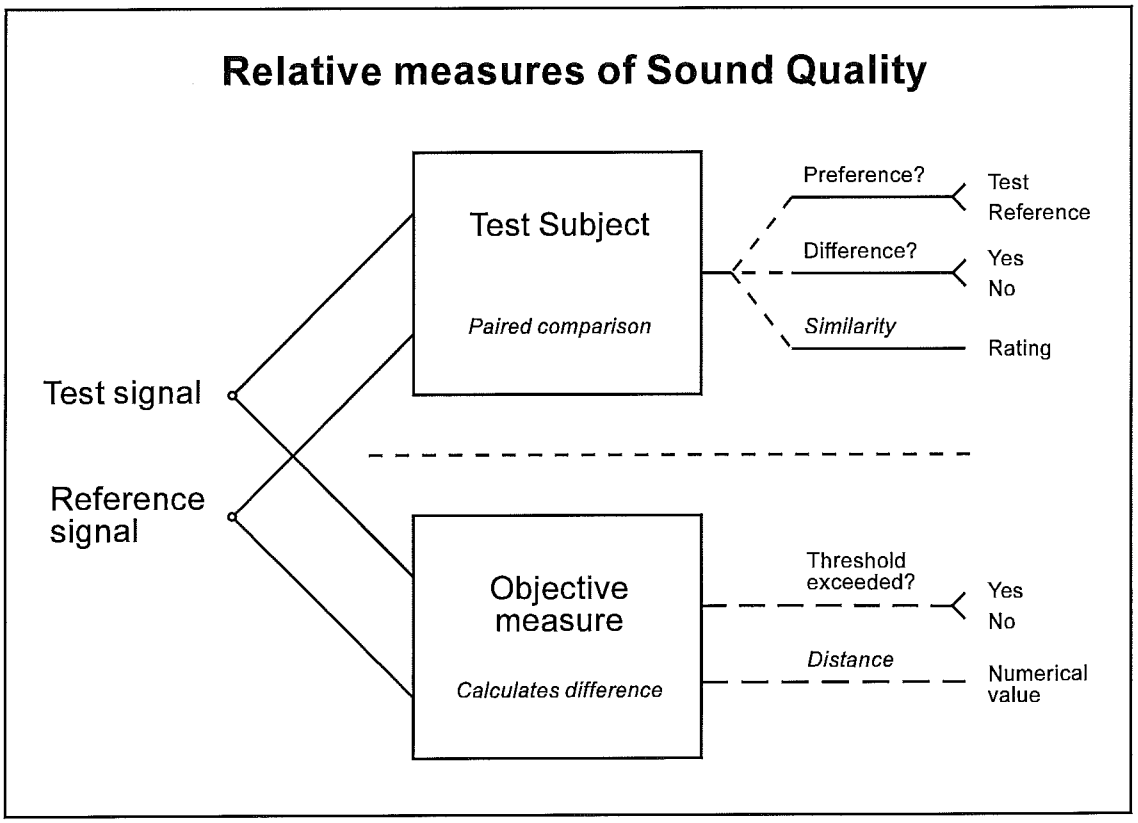


3. The CCIR recommended 5-grade impairment scale for rating of audio codecs.

The relative objective measurements are relative in the sense that they compute some kind of threshold or distance measure between two presentations of the signal. I.e. a comparison or difference of some type of signal representation is made and the outcome is related to a subjective distance or threshold measure. Even if the outcome is

numerical, it may only be applicable for a binary prediction of audibility, such as determining. whether a signal-to-noise ratio is < 0 or not.

Figure 4 depicts the types of inputs and outputs used for relative subjective and objective measures. In the following, some relative measures are discussed in the order in which they appeared in the literature.



4. Schematic representation of the inputs and outputs available to a relative Sound Quality measure. The test subject must compare the two signals and respond in one of three possible ways as shown by dashed lines. The objective measure can in principle provide both a threshold and a distance (numerical) measure.

Karjalainen (1985) presented an early application of an auditory model for prediction of sound quality of audio systems. The basic idea behind the **Auditory Spectrum Distance** (ASD) is to evaluate distortion in a psychoacoustic domain. The auditory model consists of 48 critical-band (Zwicker & Fastl, 1987) filters (Width: 1 Bark, Spacing: 0.5 Bark), implemented as 256-tap Finite Impulse Response (FIR) filters. The filterbank provides the right compromise between frequency and temporal resolution for

each band, unlike the often used Fast Fourier Transform (FFT), which has constant time resolution in all bands. This is followed by square-law rectification, addition of a base level to simulate the threshold of hearing, fast lowpass filter (3 ms) followed by a nonlinear slower lowpass filter (100 ms), simulating temporal integration and postmasking. The resulting signal is fed to a log converter to obtain the auditory spectrum in dB. This is computed for the pure sound as well as the distorted sound and the difference is calculated for each channel in each timeframe. The maximum difference across time and frequency channel is the Auditory Spectrum Distance. Detection experiments with distorted vowels indicated that ASD = 2 dB corresponded to the perceptual threshold of distortion.

Then, the quality of three vowels was rated on a 0-10 scale of Subjective Distortion Index with 1) an anchor (example of distorted vowel) fixed at 5, or 2) verbal descriptions of each scale point. The second approach is a rating scale with a fixed reference, but this is not a true similarity rating with fixed reference, since the distorted sounds are not presented in pairs with the undistorted (original) signal.

Three listeners performed the rating, and a monotonic relation between subjective distortion index and the ASD was found, with little difference between the two rating scales. There is no verification of this measure using other data or discussion of the prediction capability. It is argued in the paper that this measure outperforms traditional distortion measures by having a good correlation, but no further evidence is given. It is suggested to include some type of time-domain information in the model, like a periodicity or synchronicity measure. The model presented is a proposal for an objective sound quality measure, but the subjective ratings are so sketchy, that much further validation is required, and this has apparently never been published.

Levitt et al (1987a) presented the **Distortion Index** (DI) as a threshold measure to predict the just-detectable level of non-linear distortion. The measure is meant to be applied to hearing aids, thus it has been tested with peak clipping (output limiting) and quantization (a new type of "distortion" introduced with digital hearing aids). The

concept used in DI is to compare approximated excitation patterns (i.e. predicted masked thresholds) of the distorted (test) signal and the undistorted (reference) signal and use the maximum difference (in dB) across critical bands as the DI. The frequency gain shaping (= desired signal processing) of a hearing aid will by itself make the two excitation patterns different. Therefore, the idealized linear response has to be subtracted first in order to determine the non-linear response (= undesired) for which DI is determined. Estimating this will in practice be difficult, and the meaning of it will be unclear for a modern hearing aid with deliberately introduced non-linearities. It is a critical point, since non-linear effects on the order of 0.5 dB are to be detected. The just-detectable DI values were calculated by presenting three different vowels, distorted to threshold level. The number of subjects and subjective test method is not specified. The critical values of DI are quite different, dependent on the vowel and the type of distortion, and much work remains to make the DI a useful measure.

A very large study on objective measures of speech quality, was presented by Quackenbush et al (1988). They developed an objective, relative measure, termed **Composite Acceptability** (CA), as follows: For signals distorted, corrupted and/or coded in a large number of ways, absolute, subjective rating data were obtained and correlated to many pre-defined relative, objective signal measures of the same signals, i.e. the difference between input and output of the test object. The objective measures were fitted to the absolute, subjective ratings by means of multiple linear regression methods, and were subsequently combined to form Composite Acceptability. This "bulk" approach required enormous amounts of data, combined with massive correlation and regression analyses. With the obtained objective measure, correlations up to 0.84 with subjective data were obtained. These results are for normal-hearing listeners listening to signals that were modified in "strange" ways (i.e. speech coding) compared to the types of processing in current hearing aids, and are thus not directly applicable to hearing aid users. The data base and the various analyses from this work is very large, but the underlying objective measures suffer from the limitation that they all are pre-defined known signal measures, and that they may not be adequate for finding a

good objective measure. Also, the combination of a relative, objective measure, with an absolute, subjective measure seems conflicting.

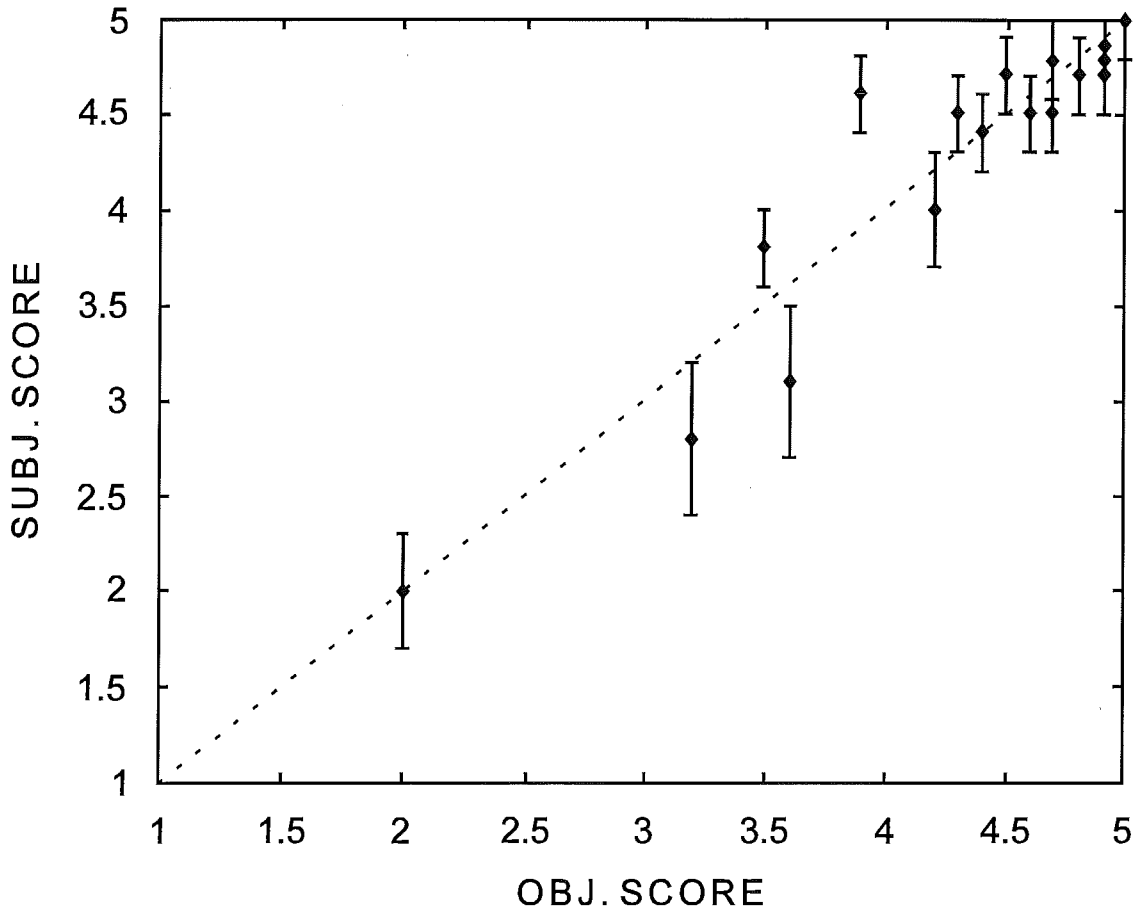
The **Noise-Masker Ratio** (NMR) was introduced in 1987 to evaluate low bit-rate perceptual codecs objectively and has been refined since then (Brandenburg and Sporer, 1992; Kapust, 1992; Herre et al, 1992; Brandenburg, 1993). An error signal is calculated as the original signal (delayed as needed) minus the coded signal, which makes it very sensitive to time alignment errors, that may not be audible. The original signal is delayed to ensure best possible alignment, by means of a cross-correlation procedure. The NMR compares the error signal to the masked threshold caused by the original signal in critical bands and expresses a margin from the noise to the masked threshold in dB.

The processing is as follows: A1024 line, 50% overlap FFT power spectrum is grouped into 27 critical bands and the excitation pattern is obtained by convolving the critical band energy with a spreading function. The critical bands and the spreading function (similar to a narrow band noise excitation pattern) are based on Zwicker's models (Zwicker & Fastl, 1990). The error signal is filtered through a critical-band filterbank without any spread of excitation and then the excitation pattern of the reference signal is subtracted band-by-band to form the local NMR. The segmental NMR can be calculated as the sum of the local NMR over the critical bands. Furthermore a "masking flag" is set if an audibility criterion is met (typically if the local NMR exceeds 0 dB) in the current time window. The NMR and masking flag can thus be considered a threshold measure, unsuited to predict subjective degradation above threshold, but providing a margin to threshold, which is useful for coder optimization.

The concept of comparing an error signal calculated directly in the time domain to the original signal makes the NMR a relative measure, however the error signal itself is never available to the listener. Because of this artificial signal, a direct subjective counterpart to the NMR does not exist (Beerends & Stemerdink, 1992). The best

equivalent is a paired comparison with the original signal as reference, i.e. a simple threshold experiment.

A numerical degradation measure (Audible Error) has later been added to the NMR concept by Kapust (1992), who also modified the NMR to use a critical-band filterbank instead of an FFT as front-end. Another interesting extension to the NMR was presented by Herre et al (1992). Although the NMR is meant as a threshold measure, it was used for prediction of a degradation scale, i.e. a numerical measure above audible degradations. The NMR in critical bands is used in a weighted sum to estimate the subjective ratings of perceptually coded signals from a previously established database (Grewin, 1993) on the CCIR 5-grade continuous impairment scale shown in Figure 3. This weighting is optimized to match the ratings by means of a Least Mean Squares algorithm. The obtained subjective results correlated well to objective scores ($r = 0.94$). The mean deviation was 0.28 (= 7%) and the maximum deviation 0.7 (= 18%). The fitting is shown in Figure 5. Only few items in the lower end of the scale were available, so the fitting of the line is based on data points with irregular spacing. No validation of the prediction scores was done with an independent data set, not previously used for the fitting (i.e. a test set).



5. NMR-calculated and subjective scores and standard deviation of the subjective grades. From Herre et al (1992).

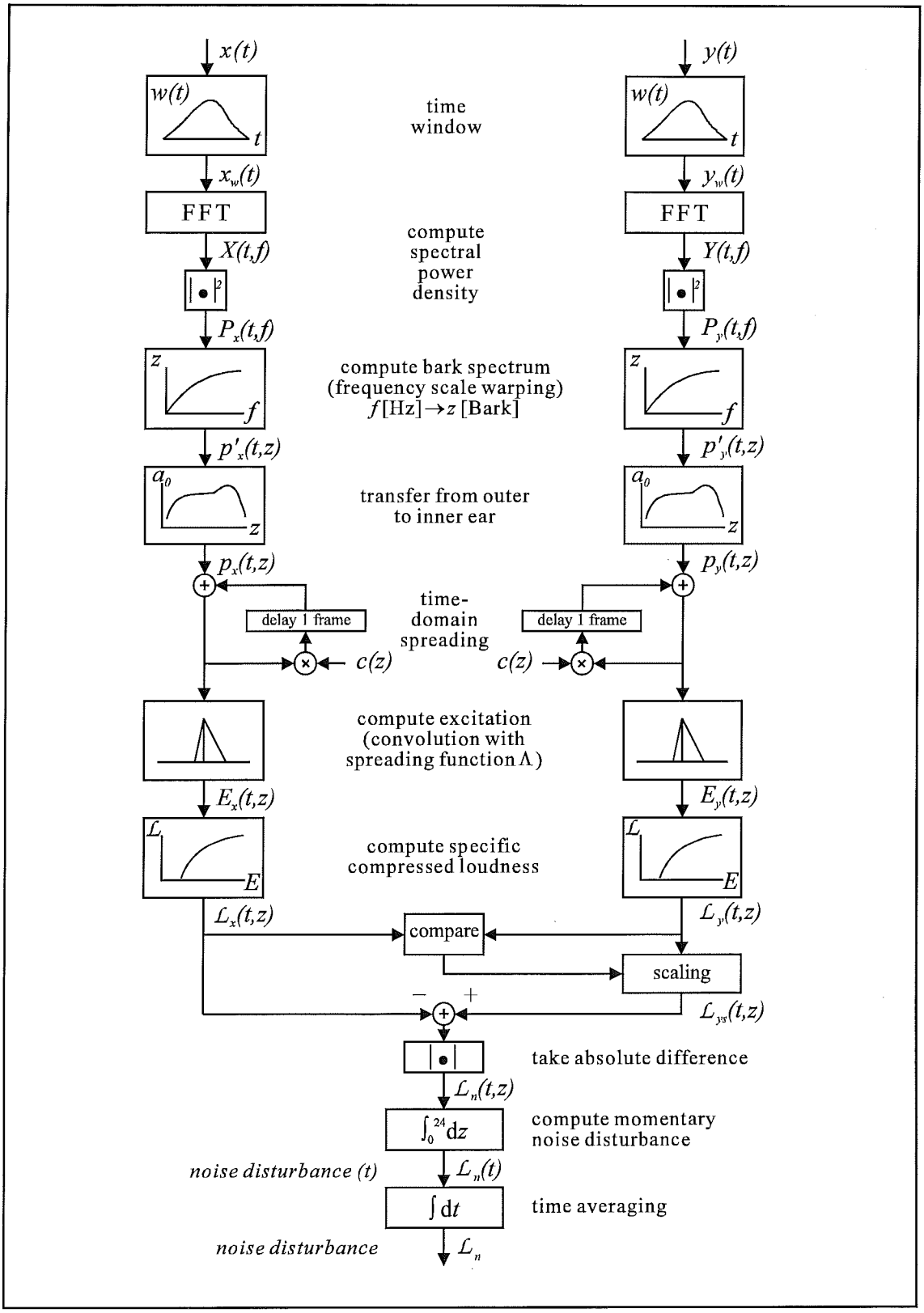
The **PERCEVAL** objective quality evaluation was proposed by Paillard et al (1992) for predicting the audibility of errors in a coded audio signal. It is a threshold measure meant for near-perfect codecs, but instead of a binary output (audible: yes or no), a probability for detection is given, which ideally must be zero. The model compares an internal representation of the original signal with an internal representation of signal + noise, and detects if there is any difference between these two, and what the probability of subjective detection is. The model calculates the error signal by subtraction in the time domain, but it is smeared in the frequency- and time-domain before adding it back to the original signal. This should make time alignment errors less critical.

The model is based on basilar membrane excitation patterns, using 2500 channels (2500 Mel - a psychoacoustic pitch scale - Zwicker & Fastl (1990)), which are derived from a

2048 pt. FFT front-end (1024-line power spectrum). In each channel, the membrane output is followed by energy-detectors with long time-constants (i.e. no synchronicity), noise is added to simulate absolute threshold and energy is log converted to form a "basilar sensation vector". Spread of excitation (masking) is also added. All this is accomplished by a matrix multiplication with the "basilar membrane energy distribution". A few implausible assumptions must be made to allow this linear transform: **1)** All mechanical processing in the inner ear is linear and time invariant (comment: most examples Report 2 contradict this), **2)** The sensitivity function on the basilar membrane is logarithmic, and **3)** The detectors along the basilar membrane have a poor time resolution (comment: other models maintain synchronicity in nervefiber responses up to about 1500 Hz (Kates, 1991)). The model simulates the common narrow-band noise as well as high-pass and low-pass noise masked thresholds (Zwicker & Fastl, 1990) with good results.

The basilar model is followed by a detection unit that uses a psychometric function centered around a Just Noticeable Difference (JND) of 2 dB for each basilar channel and the detection probabilities from all units are then multiplied (assuming independence) to form the overall detection probability. A brief experiment was done with detection of noise in various music excerpts (at constant signal-noise ratio) by five subjects - PERCEVAL predicted the detection threshold within a few dB. To be applied as an objective quality measure, much validation work on various signals and distortions is required, as pointed out by the authors.

The **Perceptual Audio Quality Measure** (PAQM: Beerends & Stemerdink, 1992) was developed to measure the subjective quality of audio devices using the concept of internal sound representation. It is a numerical distance measure that compares the internal representations of the original (reference) signal with the internal representation of the test signal. The internal representation is similar to specific loudness in critical bands (Zwicker & Fastl, 1990). A block diagram of the auditory model and the PAQM calculation is shown in Figure 6.



6. *Overview of the basic transformations used in the perceptual audio quality measure (PAQM). PAQM is calculated as logarithm of the noise disturbance. From Beerends and Stemerdink (1992).*

The reference signal, $x(t)$, and test signals, $y(t)$, are windowed and then transformed to the frequency domain. The obtained power spectra, $P_x(t,f)$ and $P_y(t,f)$, are converted from a frequency scale to a critical-band scale. Successive spectra are then passed through a low-pass filter to introduce temporal spreading (exponential smoothing). The time constant of smoothing, $c(z)$, was adjusted to fit subjective quality data. Spreading in the frequency domain is added by convolving each spectrum with a spreading function, Λ. The spectra are then passed to a power function, modeled after the calculation of specific loudness (Zwicker & Fastl, 1990). The power function exponent was adjusted to fit the subjective quality data ($\gamma = 0.04$, yielding "compressed specific loudness", compared to the usual $\gamma = 0.23$ for specific loudness). The model output is formed by taking the difference between the internal representations of the reference and the test signal, integrating it over channels, and then over time, to a single number, called the noise disturbance. The logarithm of the noise disturbance is defined as the Perceptual Audio Quality Measure (PAQM). The optimal parameter values for frequency smearing, time smearing and power law transformation were determined by fitting the model output to subjective data from a previously established database (Grewin, 1993). Fitting was better than for **PERCEVAL** and **NMR**. The fit was insensitive to changes in the temporal smearing, indicating that this is not critical for the current data reduction. A second data set of subjective quality ratings on the CCIR 1-5 impairment scale (Figure 3) from the ISO/MPEG 1991 database (Grewin, 1993) was then used for validation of the model, with good results (correlation coefficient, $r = 0.91$). There were some errors for high-quality coders, where PAQM underestimated the quality ratings by up to 1.5, compared to trained expert listeners. The authors argue that training effects should somehow be incorporated into the model. In conclusion, PAQM is not yet an acceptable predictor of sound quality, but further work is in progress. Binaural effects (release from masking) will be included in the future.

The relative measures summarized here may look for single artifacts in time-frequency and relate that to the impairment scale. They are not capable of estimating quality of linear processing, wanted or unwanted. They are suited for detection and estimation of

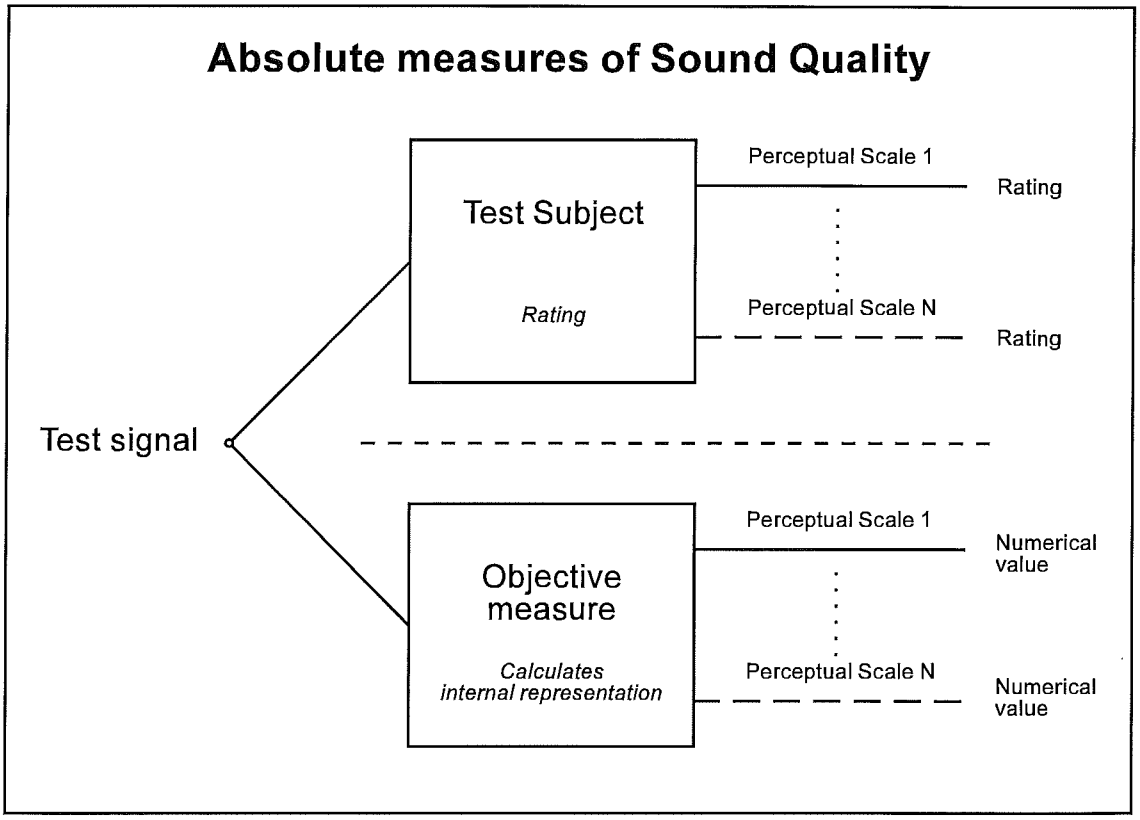
quality of detrimental ("unwanted") signal processing, whereas signal processing introduced to deliberately improve sound quality will be judged as having poorer quality by these objective, relative measures due to a perceivable difference from the reference.

2.2 Absolute sound quality measures.

Absolute quality measures do not require a reference signal for evaluation of quality, but are rated based on the test signal only. This is more representative of the end-user situation, where the user listens to a hearing aid reproducing some input signal, or to music reproduced by a hi-fi system, and judges it based on previous experience. The paired comparison situation (as used in a relative measure) is artificial, but often more sensitive. However, there may not be a reference available, no principal optimum, such as the original signal before an audio device. This is the case for a hearing-aid user - the hearing aid should not provide transparency, but provide the best possible sound quality under all circumstances. Therefore, we must rely on absolute ratings that do not distinguish between audible and inaudible processing, but between "good" and "bad" signal processing, expressed as higher or lower sound quality.

The concept of an absolute measure of sound quality is shown in Figure 7. The output from the test system - the test signal - is presented to the test subject to provide a subjective rating on one or more perceptual scales. The corresponding objective measure is normally designed to predict the ratings on the same perceptual scales, by outputting a numerical value. Neither the test subject nor the objective measure is able to completely separate the perceptual aspects of the input signal from the reproduction. However, the test subject has experience and preferences that for instance tells him or her what sound image to expect for the sound of a car.

The objective measure proposed in this report, OSSQAR, is an absolute, numerical measure.



7. Schematic representation of the inputs and outputs available to an absolute Sound Quality measure. The test subject must rate the perceptual impression on one or more scales. The objective measure should match these scales.

Much work has been done in the area of subjective ratings of sound quality, see for instance (Report 1) and (Gabrielsson and Hagerman, 1993) for a review. Within absolute measures, a large bulk of work has been done on adjective rating of sound quality, by Gabrielsson and Hagerman (1993) and co-workers. The task of the subject here, is to rate the sound quality on one or more perceptually relevant rating scales, typically by marking the rating on a 0 - 10 scale. The sound quality is usually assumed to be an underlying multi-dimensional property "within" the subject, and the choice of rating scales in an experiment may attempt to match these perceptual attributes. Experiments with a large number of perceptual attributes have been carried out and the underlying perceptual scales were then found by means of Factor Analysis (Gabrielsson and Sjögren, 1979). Examples of perceptual (also called *parametric*) rating scales are: Clearness, Sharpness, Spaciousness, Roughness etc.

Alternatively, other adjectives can be used that are more global in nature, such as Acceptability, Annoyance, Naturalness, Pleasantness, Overall Impression, etc. Such scales (termed *isometric* scales) were also used by Quackenbush et al (1988), who did a large study on the sound quality of speech coders as used in phone systems. The subjective measurements were absolute ratings of various adjectives on a rating form: Ten for the speech signal (*parametric*), seven for the background (*parametric*) and three for overall quality (*isometric*). By factor analysis (FA) these could subsequently be reduced to ten nearly independent perceptual dimensions, six for the signal and four for the background noise, which were subsequently combined into one composite acceptability measure. The rating forms and the composite measures comprised the Diagnostic Acceptability Measure (DAM). An objective measure was developed based on these ratings, termed the Composite Acceptability (CA, see section 2.1). This objective measure requires the original as well as the coded signal for comparison and is thus a relative, objective measure used to predict an absolute, subjective measure. This makes the DAM and the CA a combination of conflicting concepts.

The parametric scales usually provide more information than a single isometric scale (such as "Overall impression") about the character of the degradation/improvement (Quackenbush et al, 1988).

When using absolute scales, average ratings are obtained from a group of subjects. It is difficult for the experimenter to judge the performance of a system based on these ratings. They must be compared to ideal values, i.e. to the optimum point on each scale. For certain scales, the location of the optimum is obvious, for instance "very good" Overall Impression, the highest value on this scale is the optimum. For Sharpness, on the other hand, the optimum must lie somewhere between "very dull" and "very sharp". Gabrielsson and Hagerman (1993) have reported previous experiments (e.g. Gabrielsson et al, 1988) on rating of ideal values. At the end of the listening tests, the subjects were also asked to rate how each of the programs (music, speech) should sound in ideal reproduction. These ratings were made on the same scales as earlier, in the range 0 - 10. Some general conclusions could be made: Usually, the ideal values for Clearness,

Fullness and Spaciousness lie in the upper range of the respective scale (~ 7 - 9), while the ideal value is midway (~ 5) or somewhat higher for Brightness and Softness/Gentleness. The latter scale is roughly the inverse of the Sharpness scale, but ideal ratings of Sharpness were not made. This type of ideal ratings require posing hypothetical questions to the subjects and should thus be interpreted with caution. Gabrielsson and Hagerman (1993) also point out that rating of ideal values may seem difficult or unnatural to a hearing-impaired subject, who may not have an ideal internal reference.

Although these scales are considered absolute, some amount of **context** dependence will always exist. When presented with a certain range of stimuli, the subject is likely to adjust the use of the scales to cover a certain range. For instance, rating of telephones and ratings of high-fidelity loudspeakers should probably not be compared although they were made on the same rating scales in different experiment. Nevertheless, Gabrielsson and Hagerman (1993) state that subjects are able to develop well-defined internal references for the different scales, after due practice.

The **relation from the physical domain to the subjective domain** has been examined in several studies, but mostly on a qualitative level. For instance, Gabrielsson and Sjögren (1979) found that the perceptual scales Fullness and Brightness were influenced by the frequency response of the system. Similarly, the present experiments (Report 1) found that Sharpness was related to the amount of low-frequency energy in the signal, by high-pass filtering above 500 Hz the perceived Sharpness would increase (see also section 8).

Gabrielsson and Hagerman (1993) have summarized their experiences on the qualitative relation: *Clarity* is favored by a broad frequency range, and a slightly increasing response roughly from 500 to 4000 Hz. There will be some dependence on the incoming spectrum to the system, thus it is difficult to distinguish signal from response, although the subjects are typically asked to rate the reproduction of the system. *Fullness* is usually favored by a broad frequency response and with an emphasis on the

lower frequencies, which was confirmed in Report 1. *Brightness* increases with increased response towards higher frequencies. *Softness/Gentleness* is usually favored by a certain emphasis on lower frequencies, while its opposite, *Sharpness* is related to steeply rising frequency responses towards higher frequencies (also observed by von Bismarck (1974b)). *Spaciousness* is usually favored by a broad frequency range and possibly by a certain emphasis on mid-high to high frequencies, and it is also affected by the type of reproduction (mono/stereo, loudspeakers/headphones). *Overall quality* or *fidelity* is mostly affected by the *Clarity*, *Fullness* and *Spaciousness* scales. Report 1 confirmed that *Overall Impression* was closely related to the *Clearness* (or *Clarity*) scale with some contribution from *Fullness* as well. All the above observations were based on spectral modifications, and other non-linear side-effects of the systems evaluated (loudspeakers, telephones and hearing aids). Thus, there were no deliberate modifications of the temporal characteristics, which hypothetically might change the physical-subjective relationship outlined above.

Very little work has been found that went one step further and attempted to establish a **quantitative** relation between the two domains:

In two papers, Von Bismarck (1974a, 1974b) probably presented one of the few existing links between the physical, objective and subjective domains. In addition to the known dimensions of Loudness and Pitch, Sharpness was found to be an important perceptual attribute based on Factor Analysis of ratings on 30 rating scales. By investigation of the signal variables, Sharpness was found to increase when the upper limiting frequency or the slope of the spectrum was raised. This suggested that Sharpness is primarily determined by the frequency position of the overall energy concentration of the spectrum. The second factor was loosely found to be Compactness, which seemed to distinguish between tones and noise, i.e. between discrete and continuous spectra (von Bismarck, 1974a).

The Sharpness scale was further explored in a follow-up study (von Bismarck, 1974b), to test its consistent measurability. It was measured subjectively, by letting the subject

adjust signal parameters adaptively to produce sounds that were equal, half or double as Sharp as a reference, and this was transformed into a continuous Sharpness scale in the succeeding analysis. Furthermore, the relation between Sharpness and signal parameters was investigated. Sharpness increased with increasing lower and upper frequency limits, and with spectral slope, thus confirming the findings of the previous study. Using a different set of stimuli that varied with respect to Loudness and Pitch, it was also confirmed that Sharpness is an attribute distinguishable from both of these. Again, Sharpness was primarily related to the position of the Loudness concentration on a critical band-rate scale, rather than to a particular shape of the spectral envelope.

A calculation model for Sharpness was proposed, which integrates the first weighted moment (center of gravity) of specific loudness, N', along a critical-band scale, multiplied by a weight function, $g(z)$, equaling 1 up to 3 kHz (16 Bark) and increasing hereafter, because Sharpness increases faster than critical-band rate above 16 Bark, and divided by the total loudness of the signal:

$$S = 0.11 \frac{\int_0^{24Bark} N' g(z) z dz}{\int_0^{24Bark} N' dz} \text{ acum} \qquad \textbf{(1)}$$

This formula comes from Zwicker & Fastl (1990), based on a definition, that a narrow-band noise at 1 kHz has the absolute Sharpness of 1 acum (Latin for "Sharp"). This makes the objective Sharpness an absolute measure, although Von Bismarck (1974b) defined it as relative. It should be noted that the results from the two investigations (Von Bismarck, 1974a, 1974b) are for steady state sounds only, and that Sharpness thus may not be the dominant attribute for signals more representative of the real world, such as speech etc.

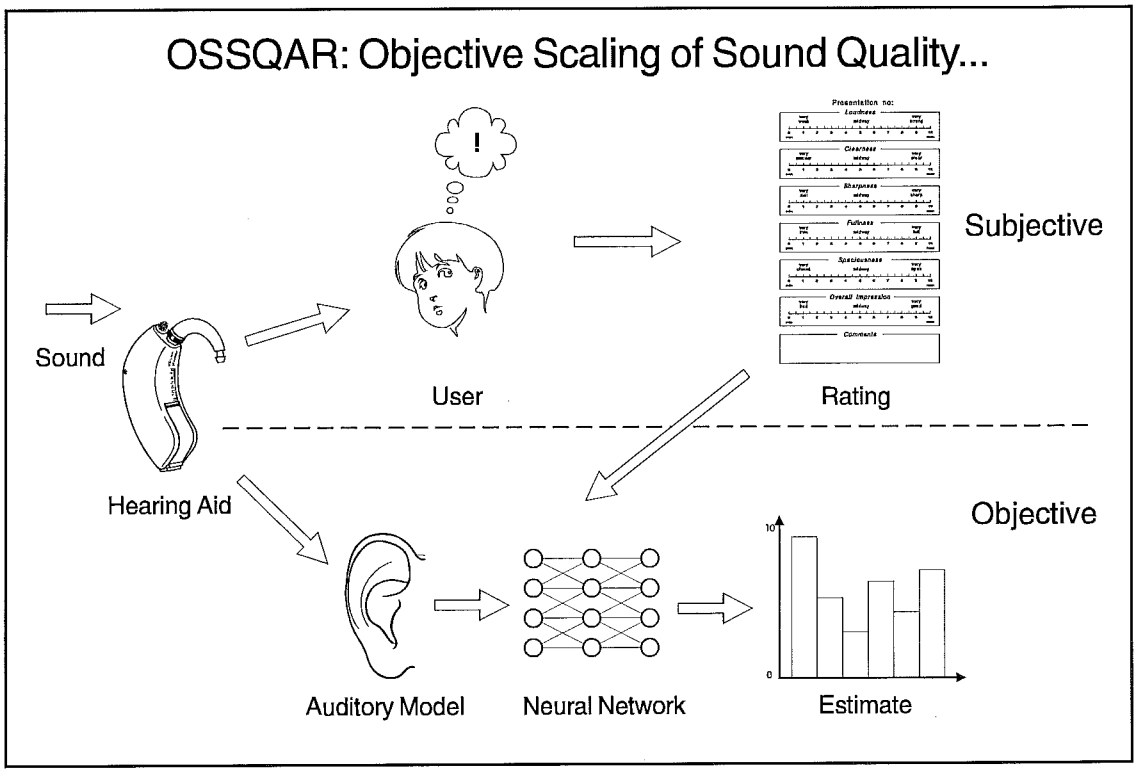
Zwicker & Fastl (1990) presented another objective, absolute sound quality measure, Sensory Pleasantness. It is based on four other known psychoacoustic measures: Loudness (sones), Sharpness (as described above), Tonality (which must be judged subjectively) and Roughness. Roughness is another perceptual measure described by

Zwicker & Fastl (1990) related to amplitude-modulation with sine waves (causing sidebands) and fluctuation of excitation within each critical band. It can be calculated for certain simple signal configurations and appears to peak at a modulation frequency of 70 Hz. A general model for calculation of Roughness is not yet available.

Sensory pleasantness was rated for a number of simple stimuli using an anchored scale, i.e. presenting a reference stimulus to the subject and dictating the corresponding rating. It is mostly affected by Sharpness, and a little for Loudness above normal conversational level. Based on these experiments a calculation model for Sensory Pleasantness has been proposed. In the present author's opinion, this measure should be used very carefully, with the limited documentation presently available.

3 OSSQAR: Project scope and goals.

The intent of the present project was to develop an objective measure of sound quality, applicable to both normal-hearing and hearing-impaired subjects. This measure and method is referred to as OSSQAR - Objective Scaling of Sound Quality And Reproduction - throughout the present report. The validity and generality of this measure should be evaluated and known limitations pointed out, so that OSSQAR is not interpreted beyond what is reasonable. With the little knowledge provided by the present literature, it was of large scientific and practical interest to find out, first of all, whether the concept of an objective sound quality measure was at all meaningful and feasible.



8. Overall project concept for OSSQAR, an objective measure of sound quality. See text for details.

The overall idea in the project is presented in Figure 8, which can be compared to the basic concept of absolute sound quality measures in Figure 7. A sound signal is processed through a hearing aid, where it is subject to "desired" signal processing

(frequency shaping, compression, output limiting etc.) and "undesired" signal processing (resonance peaks, non-linear distortion etc.). The subjects in the present experiment - one normal-hearing group and one hearing-impaired group - must rate the subjective sound quality of the signals on a number of perceptual scales (e.g. Loudness, Clearness, Sharpness). The same signals are presented to a computer model of the ear, which mimics some of the important, presently known psychoacoustic properties of the normal or the impaired ear.

In the actual rating experiment, the input signals were not processed through a hearing aid, but through various types of signal operations: addition of background noise, filtering, clipping, compression etc., in various combinations (see section 8.1 and Report 1 for details). The purpose of this scheme was to generate signals that were perceptually very diverse and elicited responses also on the extremes of the different perceptual dimensions.

To verify this idea and study some of the underlying assumptions and relations, a number of sub-tasks, of both research and development character, were formulated at the outset of the project. These are described in detail in the four research reports, that document the entire project:

1. To find and evaluate subjective methods for evaluation of sound quality, and study the reliability and range of the quality judgments for both normal-hearing and hearing-impaired subjects. (This Report, Section 4 and Report 1).

2. To investigate the influence from various experimental parameters on perceived sound quality: stimuli (test signals), hearing impairment, signal and processing parameters (i.e. input signal and hearing aid). (This Report, Sections 4 and 8 and Report 1).

3. To study the properties of the rating scales: Are they absolute? Are there redundant scales? Is the interpretation the same for both normal-hearing and hearing-impaired groups. (This Report, Section 4 and Report 1).

4. To develop and evaluate a computer software model of the external, middle and inner ear, based on either the physiology (a cochlear model) or the psychophysics (an auditory model) of the ear. The model should

include the impairments associated with sensorineural hearing loss. The rationale for using an auditory model is that the internal representation of sounds is the appropriate one for evaluation of sound quality. (This Report, Section 5 and Report 2).

5. To link the subjective quality ratings with the ear model by means of an artificial neural network. This is an attractive way to approximate the multi-dimensional, non-linear mapping from the ear model to the subjective sound quality. (This Report, Section 6 and Report 3).

6. To assess the benefits and limitations of OSSQAR. How much prediction error can be expected? Is OSSQAR truly an absolute measure? What are the optimum values of OSSQAR? (This Report, Section 7).

7. To study the obtained mapping from the physical world to the perception of sound quality. (This Report, Section 8 and Report 1 and Report 3).

The division of the project into sub-tasks and separate reports also ensured that the many useful general results were published, independent of the outcome of OSSQAR.

The proposed measure is new in a number of ways, compared to the other suggested measures from the literature (section 2): It is designed to be an absolute measure, which is required when an ideal transparent system is not available for comparison. This is the case for a hearing aid, where deliberate modifications to the sound signal take place. Furthermore, OSSQAR predicts perceptual scales, rather than "global" scales, such as Naturalness or Degradation, and it was based on subjective measures designed and obtained specifically for this purpose. OSSQAR is also the first example of an objective sound quality measure applied to the hearing-impaired population.

4 OSSQAR: Subjective measures.

The subjective evaluation of sound quality was the first step in developing OSSQAR and is fully described in Report 1. The study is summarized below:

4.1 Purpose

A major goal of the study was to obtain quantitative, reliable sound quality ratings for the development of OSSQAR. A shortcoming of previous studies was the limited use of hearing-impaired subjects for quality assessment of hearing aids. It has been argued that normal-hearing subjects are representative of the hearing-impaired, but more sensitive to small changes in the signal. This assumption is not obvious, when the dramatic changes in frequency and temporal resolution due to hearing impairment are considered. It was thus decided to put equal emphasis on hearing-impaired subjects, and to verify whether the two subject groups had identical interpretations of the rating scales. Furthermore, the investigation was to include a large number of diverse signal processing conditions in an attempt to obtain general results and to make the subjects use a wider range on each perceptual scale.

4.2 Materials and methods.

Subjects. A total of 12 normal-hearing (NH) and 11 hearing-impaired (HI) subjects were used. The NH subjects were selected to have pure-tone thresholds less than 15 dB HL across all audiometric frequencies. The NH group consisted of 8 females (age 19 - 34) and 4 males (age 19 - 30). The HI subjects were selected to match a pre-defined sensorineural hearing loss, typical for the average hearing-aid user. By restricting the hearing loss to a narrow range it was possible to use the same amplification for all HI subjects. The HI group consisted of 6 females (age 32 - 80) and 5 males (age 64 - 83) who were all hearing-aid users. All subjects, except for two from the NH group, had previous experience as research subjects.

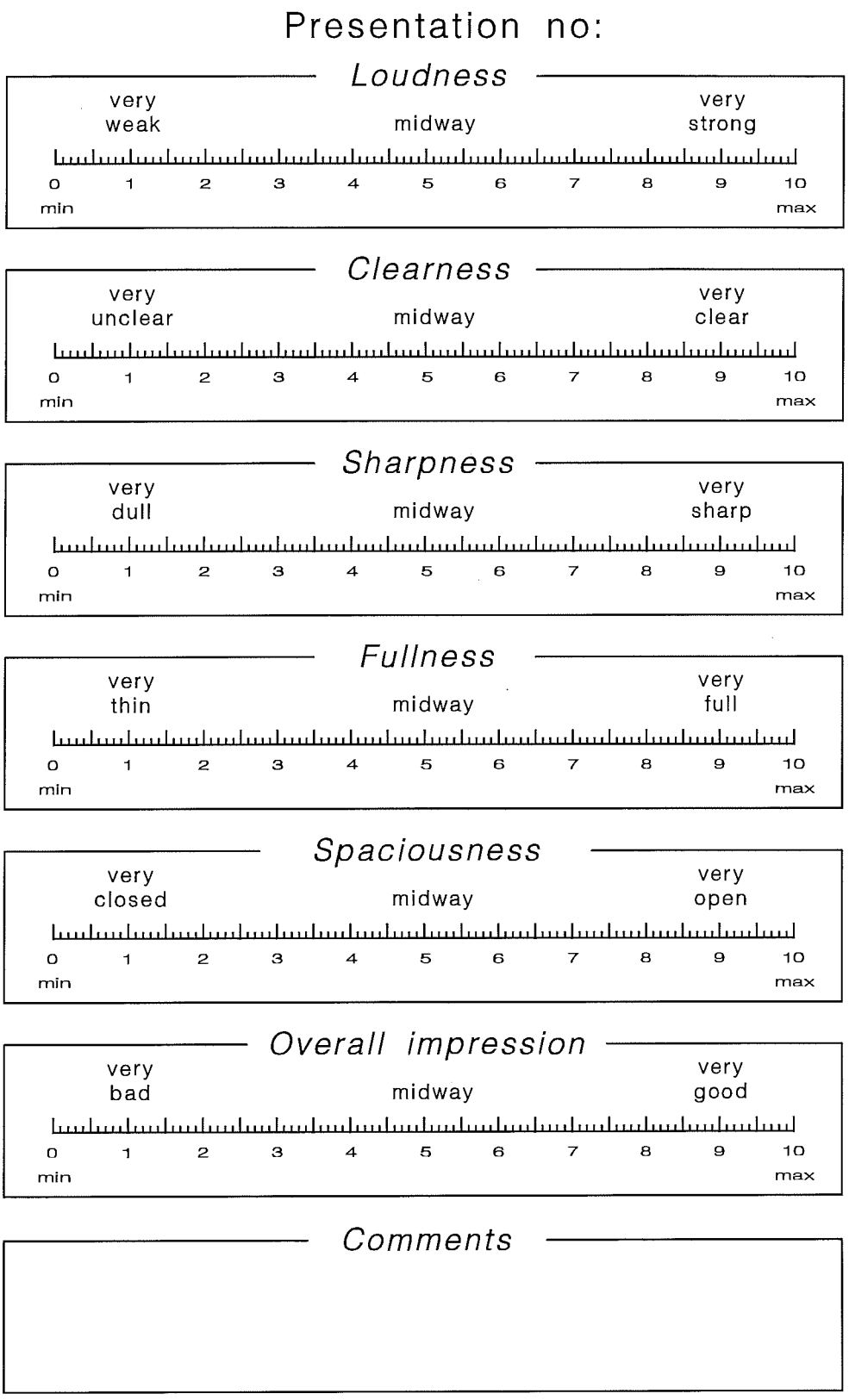
Signals and apparatus. The experiment used two clean signals in combination with a noise signal as input to a signal processing structure. The variable factors in this structure provided for 256 combinations, from which 64 were picked to form a fractional factorial experiment (Box et al, 1978). The signal processing options included signals with or without added background noise (+5 dB signal/noise ratio (S/N) for speech, + 10 dB S/N for music) split into three frequency bands and subsequently passed through, compressed, clipped or removed before being summed to form the stimulus. The clipping was set to the 50% point (L_{50}) based on the amplitude distribution function of the signal in each band. The full-range amplitude compression (limiting) used a 1:20 compression ratio, knee point set 20 dB below L_{50}, and attack and release times of 20 and 200 ms, respectively. The 64 stimulus files for the normal-hearing group were multiplied by individual scale factors to equalize the long-term level (L_{eq}), in order to keep the perceived loudness approximately constant. After scaling, 64 new stimulus files for the hearing-impaired group were generated, by convolving with a digital filter, providing the proper frequency-dependent amplification according to the POGO II rule (Schwartz et al, 1988). All stimulus files were 30 sec. in duration, and always played twice in succession, allowing the subject one minute to rate each stimulus.

The signal files were played from the hard-disk of a PC, using the Ariel DSP-16 signal processing board as a digital-to-analog converter, followed by a 10 kHz low-pass filter. A manual attenuator was used to set the signal level to Most Comfortable Level (MCL) once for each subject. The signal was delivered monaurally to the best ear of the subject via Sennheiser HD250 Linear II headphones. All listening took place in a sound-proof audiometry test booth.

Rating scales. The rating scales and rating procedure were based on previous work (Gabrielsson et al., 1988), from which six perceptual scales were chosen:

- Loudness
- Clearness
- Sharpness
- Fullness
- Spaciousness
- Overall Impression

The scale was designed as a horizontal line with numerical markers and verbal labels for the midpoint and the two extremes of each scale. The rating form (English version) is shown in Figure 9.



9. *English translation of the Danish rating form used in the rating experiment (from Report 1).*

Before each session, the subject received a short written instruction on how to perform the rating task, plus a brief written description of the midpoint and the two extremes of each of the six rating scales.

Rating procedure. After audiometric control and interview, each subject participated in three rating sessions, one per day. At the beginning of the first session, the overall stimulus level was set by means of an up-down MCL (Most Comfortable Level) procedure, to ensure a comfortable stimulus level. During each session, 5 * 16 stimuli were rated, i.e. 16 stimuli for warm-up (these data were discarded) plus all 64 stimuli in a pre-defined randomized order. Each block of 16 stimuli was followed by a short break. Thus, the entire experiment was repeated three times for each subject.

4.3 Main results.

The data from the rating forms were entered into a spreadsheet for further statistical analysis. No transformation or normalization was applied to the data, and it was assumed that the rating scale data followed a normal distribution. No strong theoretical arguments against this could be raised. A two-way analysis of variance (ANOVA) was applied to each subject for each rating scale. This was done to ensure that each subject was reliable and useful in the group analysis. It was found that all 12 normal-hearing (NH) subjects had significant stimulus effects on all six scales ($p < 0.01$), except one subject on the Loudness scale. All NH subjects had significant day-to-day changes on one or more scales ($p < 0.05$). For the 11 hearing-impaired (HI) subjects, all had significant stimulus effects on all six scales ($p < 0.01$), except one subject on the Sharpness scale.

Effects of signals and subjects. In order to make general statements concerning the subject populations, all subjects were included in six analyses of variance (ANOVA), one for each rating scale. These ANOVA's contained the four main effects: Stimulus, Group (NH vs. HI), Subject (within group) and Day. One NH subject was left out of the analysis to balance the design (thus 11 subjects in both groups). The mean squares

(MS) of these effects have been normalized according to the number of conditions for each (i.e. the stimulus MS was divided by 64), to indicate the average change on the scale for a change in stimulus, subject, etc. These normalized expected mean squares (EMS) are listed in Figure 10 for each of the six scales.

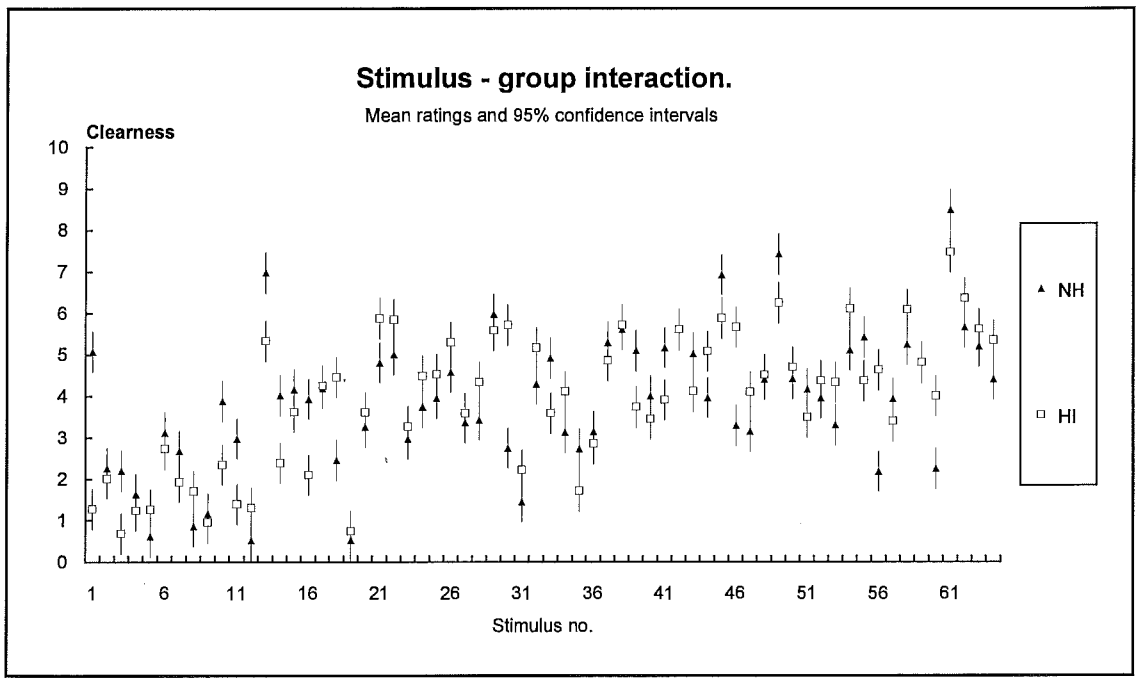
Effect size (from EMS analysis)

Rating scale:	Loudness	Clearness	Sharpness	Fullness	Spacious.	Overall
MAIN EFFECTS						
Stimulus (S)	**0.13**	**2.03**	**1.31**	**1.19**	**0.95**	**2.16**
Group (G)	0	0	0	0	0	0
Subject (P)	**0.24**	**0.42**	**0.43**	**0.4**	**0.59**	**0.46**
Day (D)	0	0.02	0	0	0	0.03
INTERACTIONS						
Stimulus - Group (SG)	**0.32**	**0.63**	**0.31**	**0.31**	**0.37**	**0.35**
Group - Day (GD)	0	0	0	0	0	0
Subject - Day (PD)	**0.03**	**0.06**	**0.06**	**0.06**	**0.1**	**0.08**
RESIDUAL	0.9	2.08	2.23	2.1	2.5	1.57

10. *Summary of ANOVA for all subjects. Numbers shown are normalized mean squares of effects (see text). Bold type indicates a significant effect ($p < 0.05$). From Report 1.*

The significant main effects ($p < 0.05$) are indicated as bold numbers. This shows that the 64 stimuli were different on all scales, but with a relatively smaller effect on the Spaciousness and Loudness scales. The small effect on Spaciousness is not surprising, considering that all stimuli were presented monaurally in headphones. The small effect on Loudness confirms that Loudness was kept almost constant across the 64 stimuli. There was no difference between the two groups (NH vs. HI), meaning that one group is not shifted left or right on the rating scale compared to the other. There is a significant difference between Subjects, i.e. the subjects use the scales differently, but the subject effect is numerically smaller than the stimulus effect. There is no overall difference from day-to-day, i.e. no systematic shift left or right on the rating scales during consecutive sessions.

In the experimental design used here, it is also possible to examine certain interactions, as listed in Figure 10. The significant Stimulus-Group interaction indicates that the two groups (NH and HI) disagree on the rating of the stimuli - this can be due to the difference in hearing capacity and/or the age difference. Given this fact, we must conclude that hearing-impaired and normal-hearing subjects cannot be equated in the present and future experiments. There was furthermore a significant Subject-Day interaction, meaning that the subjects had different training / adaptation patterns. The Stimulus-Group interaction can be visualized by plotting all stimulus means separately for the two groups, as shown in Figure 11 for the Clearness scale.



11. Mean ratings of Clearness for the 64 stimuli with 95% confidence intervals. The normal-hearing group is represented by filled triangles (NH) and the hearing-impaired group by open squares (HI). From Report 1.

The graph shows that the experimental design elicited responses over a broad range on the scale, with stimulus means covering almost the entire range (0 - 10). The majority of responses are under mid-way, meaning rather low Clearness in general. The Stimulus-Group interaction is evident in the graph from the non-parallel course of the curves for the two groups. Generally, the means for the NH group are more spread out

on the scale, i.e. the NH group uses a wider range on the scale. This can also be interpreted as a higher sensitivity for the NH group.

Signal and processing effects. Since the stimuli were generated as a systematic combination of signals, noise and processing options in a factorial experiment, the effects of these parameters on each rating scale could be estimated and tested for significance. This is summarized in section 8.1, where the qualitative relation between stimulus parameters and perceived sound quality is discussed.

Rating scales and perceptual dimensions. Three questions were addressed concerning the properties of the rating scales: **1)** To what extent are they correlated, **2)** which of them are relevant in describing the perceived sound quality adequately, and **3)** are they interpreted the same way by the two subject groups?

The correlation matrix is given in Figure 12, showing the normal-hearing group above the diagonal and the hearing-impaired group below the diagonal.

	Loudness	Clearness	Sharpness	Fullness	Spacious-ness	Overall Impress.
Loudness		0.29	0.1	0.1	0.16	0.23
Clearness	0.44		-0.15	0.48	0.42	**0.84**
Sharpness	0.34	-0.01		-0.47	-0.15	-0.29
Fullness	0.31	**0.59**	-0.27		0.4	**0.6**
Spacious.	0.44	0.49	0.15	0.37		**0.53**
Overall	0.38	**0.83**	-0.05	**0.64**	**0.5**	

Normal-hearing subjects

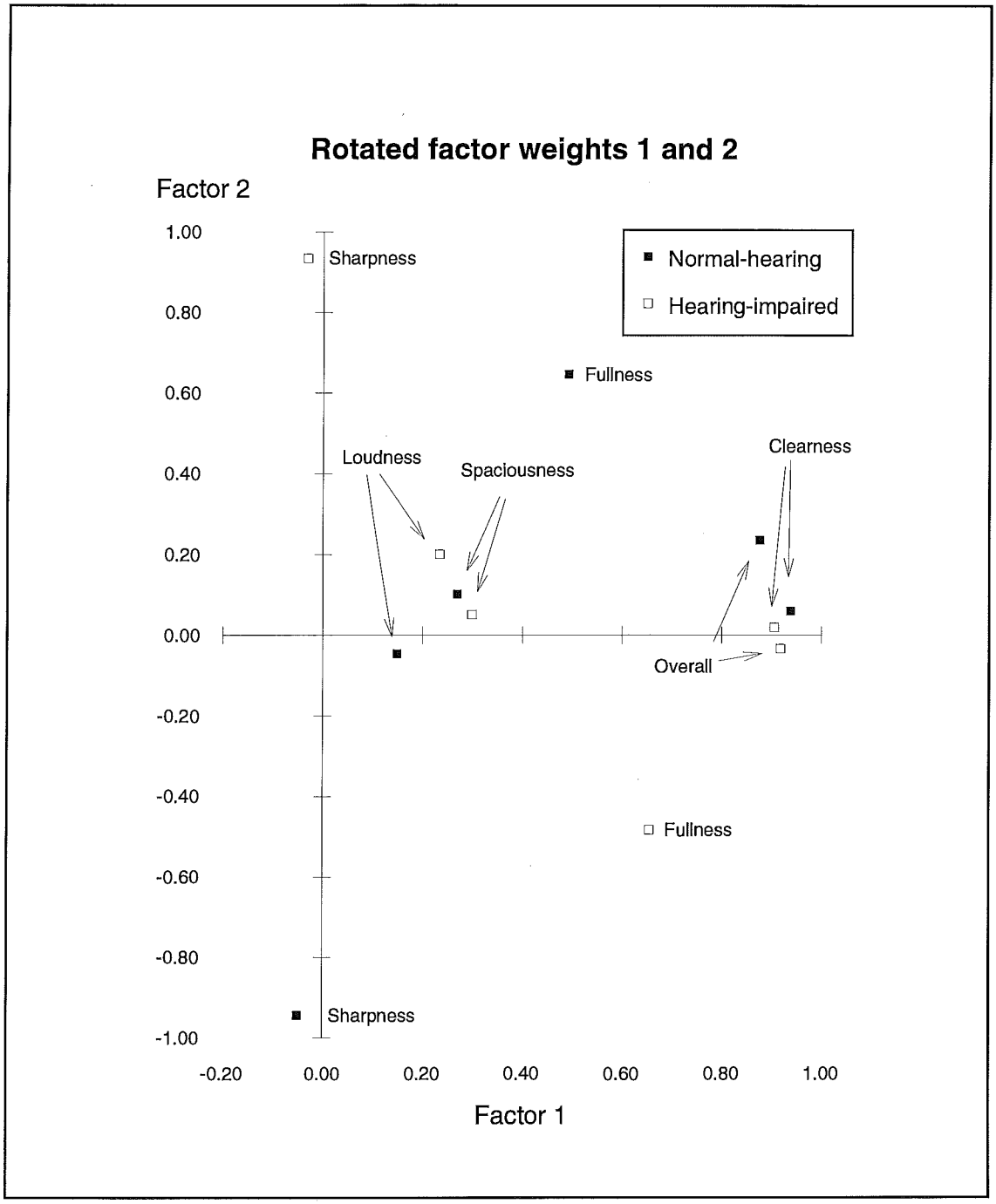
Hearing-impaired subjects

12. Correlation matrix of the rating scales for the two subject groups separately. Correlation coefficients ≥ 0.5 are in bold types. From Report 1.

All intercorrelations are significant ($p < 0.0001$), due to the large number of observations, but not necessarily meaningful. Using a criterion of $r \geq 0.5$, a few scales can be considered important correlates:

- Overall impression and Clearness for both groups.
- Overall impression and Fullness for both groups.
- Overall impression and Spaciousness for both groups.
- Fullness and Clearness for the hearing-impaired group.

The remaining scales are poorly correlated, indicating that more than one of the scales is necessary to adequately describe the sound quality. This was analyzed further in a factor analysis, where the underlying perceptual dimensions can be derived from the correlation matrix. For the normal-hearing group, *90.8%* of the data variance was accounted for by four factors. For the hearing-impaired group, four factors accounted for *91.4%* of the total variance. These four factors were extracted and subjected to a VARIMAX rotation (Ferguson and Takane, 1989), in order to align a coordinate system with the principal components in the data. The rotated factor weights of the original six scales is the position of the six rating scales in the new, four-dimensional factor space. This is shown for the primary two factors in Figure 13.



13. Rotated factor weights 1 and 2 for the two subject groups. For both groups, factor 1 accounts for roughly 50% of the total variance and factor 2 accounts for roughly 21%. From Report 1.

The factors are ordered in descending order of contribution to the total variance. **Factor 1** accounts for 47.9 % (NH) and 50.7 % (HI) and can be interpreted the same way by the two subject groups: It is dominated by equal contributions from Clearness and Overall impression with some contribution from Fullness. This shows that Overall Impression

is a redundant scale, equally well covered by Clearness. **Factor 2** accounts for 20.2 % (NH) and 22.7 % (HI) of the total variance and it is dominated by Sharpness and Fullness, the two having an opposite effect. The two subject groups have opposite orientation along Factor 2, due to slight differences in the factor analysis, however the perceptual interpretation is the same. Factor 2 may be interpreted as low-frequency vs. high-frequency spectral content, i.e. a low-frequency dominated stimulus will be rated very Full and very Dull (not Sharp), and opposite when much high-frequency energy is present. Sharpness and Clearness were also found to be the most significant scales by Gabrielsson and Sjögren (1979), but in the opposite order of importance compared to the present study. However, these earlier studies mostly used spectral modifications of the signal, with no added background noise and no explicit temporal processing.

Factor 3 accounts for 12.4 % (NH) and 9.3 % (HI) of the total variance and it is dominated by Spaciousness (NH) and Loudness (HI). Similarly, **Factor 4** accounts for 10.4 % (NH) and 8.7 % (HI) of the total variance and it is dominated by Loudness (NH) and Spaciousness (HI). The two scales are thus in opposite order for the two groups. The low importance of Loudness was expected, since it was attempted to keep Loudness constant in the experiment, although less successful for the HI group.

In the present factor analysis, there is very good agreement between the two groups with respect to correlation between scales and the location of the rating scales in the underlying factor space, thus we can conclude that normal-hearing and hearing-impaired listeners perceive sound quality in the same perceptual space, and both groups use the same interpretation of the scales.

4.4 Conclusion: Subjective sound quality.

The sound quality rating experiment had a number of important outcomes:

All subjects performed the rating task reliably, and could distinguish the stimuli, according to statistical analysis. The rating data covered a wide range on each scale,

which is important for the development the present objective sound quality measure, OSSQAR.

The two subject groups (normal-hearing vs. typical sloping hearing loss) did not differ in mean ratings on any of the scales. Assuming that the auditory perception is not identical for the two subject groups with very different hearing configurations, it may be concluded that the rating scales are not absolute. The normal-hearing group used a wider range on the scales, and can be considered more sensitive.

The perceived sound quality can be described by four underlying dimensions, with two dominant scales: **1)** Clearness combined with Overall Impression, and **2)** Sharpness and Fullness. The two subject groups appeared to interpret the rating scales identically, thus Sharpness and Clearness are the same perceptual attributes for both groups.

5 OSSQAR: Auditory modeling.

The specification, implementation and evaluation of an auditory model with hearing loss was the second step in developing OSSQAR and is fully described in Report 2. The report is summarized below:

5.1 Purpose and motivation.

The overall project idea (section 3) calls for an auditory model to be used as pre-processor for the artificial neural network. It is not clear whether this type of advanced signal processing is necessary for a good objective estimate of sound quality. One could imagine using 1/3 octave filtering, short-term FFT analysis or other signal measures with more or less perceptual relevance, assuming that the neural network would still be able to establish the underlying relation between signal and perceived sound quality. Nonetheless, the present investigation used a perceptual model of the ear as a signal preprocessor, which includes known properties of the normal and impaired ear (Report 2). Relevant, similar applications have been found in speech recognition literature (Report 3), where auditory-based front-ends often appear to provide a benefit compared to more traditional pre-processing (such as FFT analysis).

Other proposed objective measures of sound quality are usually also based on some kind of perceptual model that calculates some kind of "internal" representation, i.e. a psychoacoustic property, such as specific loudness. The sound quality is then predicted based on this internal representation or comparisons of two different internal representations (see section 2).

In the neural network literature, it is often emphasized that "everything we know" should be processed intelligently, and only the rest "that we don't know" should then be left for the neural network to solve. Based on this philosophy, the present auditory model should implement the known features of the peripheral hearing system, while leaving the more "central" part to the neural network. Since the present project is

specifically concerned with the perceived sound quality for hearing-impaired listeners, the auditory model should also include the known changes to the auditory system and their dependence on hearing loss.

The present auditory model was developed as preprocessor for the current project, but with other applications in mind. The psychoacoustic principles presented in the literature are mostly based on simple stimuli, such as pure tones and broad-band noise. By implementing these in a practical signal processing program, it is possible to study the perception of real sounds (e.g. speech) including the complex nonlinear interactions in the auditory model. The present auditory model can thus calculate masking patterns and specific loudness in auditory filter bands, including the calculation of total loudness in both the normal and the impaired case, according to established theories. There are major similarities between the present model and a recently published model for estimation specific loudness in relation to hearing-aid fitting (Leijon, 1989; Leijon, 1990).

A model of hearing can be based primarily on either the physiology of the inner ear or on the psychophysics of hearing. Several physiological models (cochlear models) were reviewed in Report 2. These models either simulate the micro- and macromechanics and neural functions of the cochlea or attempt to match physiological data (e.g. single nervefiber tuning curves) by means of filters, nonlinear processing etc. A major problem with cochlear models is the lack of consistent, reliable quantitative data on tuning curves etc., for humans with typical sensorineural hearing losses. In fact, all physiological data on impaired hearing are based on losses induced either by severe deliberate noise-exposure or by ototoxic drugs. Cochlear models are however still of interest for more qualitative research studies to further our knowledge of hearing physiology. Several such have been described in the literature (Allen, 1985; Seneff, 1985; Lyon, 1982 and Kates, 1991).

Psychophysically based (auditory) models are based on behavioral measures of hearing, i.e. thresholds, masking curves, loudness growth etc. These psychoacoustic properties

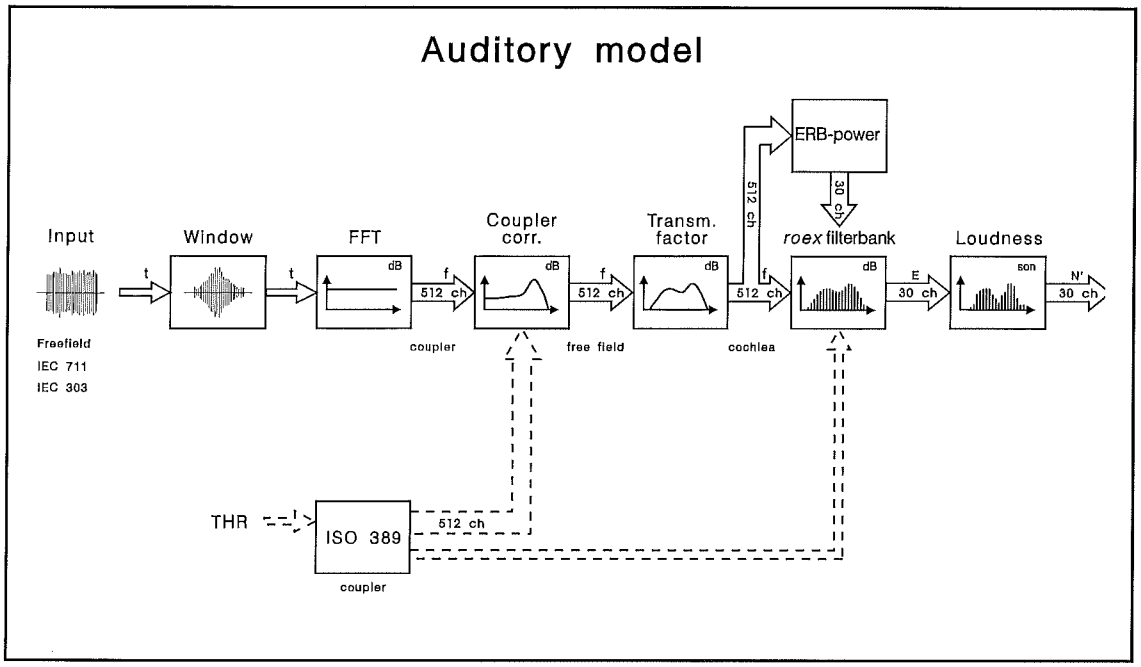
have been measured extensively on subjects with normal hearing and with sensorineural hearing loss, and are available in the literature. Previous examples of auditory models based on psychophysics of normal hearing have been presented by (Cohen, 1989; Hermansky, 1990 and Karjalainen, 1984, 1985). The only example of an auditory model with hearing loss was presented by Leijon (1989, 1990).

The shape of the auditory filters have been modeled and measured for both normal-hearing and hearing-impaired subjects (Glasberg and Moore, 1986; Glasberg and Moore, 1990), in a manner that made it very tractable for implementation in a practical, quantitative model of the normal and the impaired ear.

Thus, the present auditory model was based on psychoacoustics alone. The same choice was made by Leijon (1989). The present model is summarized below.

5.2 Model description.

The processing steps in the model are outlined in Figure 14. This can be compared to the similar model used for PAQM (Perceptual Audio Quality Measure) shown in Figure 6 (Beerends & Stemerdink, 1992).



14. *Block diagram of the auditory model. Solid lines indicate signal paths, dashed lines indicate control parameters (threshold parameters for loudness initialization and to control the filter bank). See text for details. From Report 2.*

The core of the model is a set of filters shaped as rounded exponentials (*roex*) in the frequency domain. These filter shapes are derived from detection thresholds of pure tones masked by notched noise, with the tone located in the notch (Glasberg & Moore, 1990). Contrary to the original critical bands that were specified in terms of cut-off frequencies only (Zwicker & Feldtkeller, 1967), the filter shape is specified, and the output of the filterbank output is the excitation pattern (E), which includes frequency masking effects automatically, by virtue of the sloping filter shapes. In models based on critical bands, the filter bank uses rectangular bands and the excitation pattern must be calculated afterwards by convolving, in the frequency domain, with a *spreading function*, similar to a narrow band masking pattern (Brandenburg, 1993).

The present model encodes loudness, according to the models by Zwicker & Feldtkeller (1967) and Zwicker & Fastl (1990): Specific loudness (N') is calculated from the excitation in each critical band (here: each filter channel), by means of a power function with exponent 0.23. The threshold of hearing is considered equal to an internal masking

noise, hence there is a steep growth of loudness close to threshold. The total loudness can be calculated by summing the specific loudness across all bands.

The present auditory model performs the following operations on the signal:

- The incoming signal (t) is windowed to a user-specified frame-size.
- An FFT analysis is performed on the windowed signal and a power spectrum (f) is obtained.
- An equalization is then applied to the power spectrum to compensate for the frequency response of the coupler, in which the signal was recorded.
- In the same way, a transmission factor is applied by multiplication in the frequency domain. This factor can be interpreted as the linear transmission characteristics of the ear canal and the middle ear.
- The signal power is determined in rectangular bands (or wider, in the hearing-impaired case), by summing the power spectrum (f) within the limits of each band. These power values are used to adjust the filterbank:
- The resulting power spectrum is then passed through a filterbank, consisting of 30 auditory *roex* filters whose shape depend on hearing loss and on the signal power. The *roex* filterbank output is the excitation pattern (E).
- The parameters for hearing loss (THR) are converted from dB Hearing Level (HL) to dB Sound Pressure Level (SPL) and used to influence frequency selectivity in the filterbank and sensitivity in the loudness function. These initialization parameters are indicated by dashed lines.
- The *roex* filterbank output (E) is passed on to the specific loudness function that converts excitation in each channel to specific loudness, (N'). The absolute threshold of the subject is taken into account here. N' is the default model output, but the output can be taken at other points in the model. The total loudness of an incoming signal can be calculated by summing the specific loudness across bands.

No temporal effects have been incorporated into the model presently, neither for normal-hearing (NH) nor hearing-impaired (HI) listeners. Temporal integration and post-masking are factors that determine the temporal resolution of the human ear, and these may be affected by hearing loss.

In the current configuration the auditory model represents a combination of different "schools" and experimental results from the psychoacoustical literature. In an attempt to create a coherent, practical and useful model, many compromises have been made. The psychoacoustical literature contains disparate results and focuses on separate aspects of hearing, one model will thus not be able to unite all these results in a meaningful way. Given the large variance in research conclusions and many remaining unanswered questions, the model results should be interpreted with caution.

The auditory model has been implemented as a PC program that reads waveform signal files and outputs the results to different optional file formats. Various parameters for the auditory model, including hearing loss, are specified in an accompanying parameter file. With the parameters for sample rate, FFT length etc., used in the present project, the model could process the signals in roughly 60 * real time on a 486/25 MHz PC. See Report 2 for details.

5.3 Verification.

A number of classical psychoacoustic tests were imitated with the auditory model and the simulation results were compared to experimental data from the literature:

- Pure tone masked thresholds as a function of signal level and frequency. Data on NH listeners were taken from Zwicker & Fastl (1990). The model predicted the masked thresholds well, except at very high levels, where the upward spread of masking was underestimated.

- Masked thresholds for white noise - NH listeners (Zwicker & Fastl, 1990). Some disagreement at lower frequencies. This is the region where the two auditory filter models - critical bands (Zwicker & Feldtkeller, 1967) and *roex* filters (Glasberg & Moore, 1990) disagree the most.

- Narrow-band masked thresholds for subjects with mild-to-moderate hearing losses. Data was taken from Dubno & Schafer (1992). There was good agreement between predicted and actual thresholds.

- Growth of loudness for pure tones, NH listeners. There was good agreement with actual loudness ratings (Scharf, 1978), except at high levels. There, the model underestimated loudness, probably due to the upward spread of excitation limit at 7 kHz (the center frequency of the highest channel).

- For hearing-impaired listeners, little experimental data was available. Compared to one study (Hellman and Meiselman, 1990), that fitted a different loudness growth model to magnitude estimations of loudness, the auditory model had a steeper initial growth of loudness and some overestimation at medium levels. At high levels, where all loudness models predict normal loudness perception, there was reasonable agreement.

As an example of a real-world application of the model, loudness spectrograms for a speech utterance were presented. By introducing hearing loss, the speech sounds became less audible and less detailed, a problem that linear amplification did not compensate properly. This demonstrated how the model could be used for hearing aid development and evaluation (Report 2).

5.4 Conclusion: Auditory model.

An auditory model based on psychoacoustic theory was presented. The advantage of this approach, rather than using a physiological model, was discussed. The model has been specified, developed and implemented based on selected results from the literature.

The elements of the model are: Power spectrum calculation, equalizations and coupler corrections, an auditory filter bank with or without hearing loss, and loudness growth functions for normal and impaired hearing. The temporal properties of the normal and impaired hearing system have not been included in the current implementation.

The model was verified against various results from the psychoacoustic literature. For normal hearing, the model reproduced masking patterns for narrow-band noise well, underestimating upward spread of masking at high masker levels. For hearing-impaired subjects, the upward spread of masking was furthermore limited by the upper frequency

limit used for the simulations. Nevertheless, the model correctly reproduced narrow-band masked thresholds well for a small, selected group of impaired subjects.

The loudness growth function was generally correct, but loudness was underestimated at high levels, compared to the usual 0.3 power law used at high levels. This discrepancy was also due to the frequency limit in the model. As a consequence, the equal loudness level contours for normal hearing were also incorrect at high levels. For impaired hearing, the model also produced the proper loudness growth according to Zwicker & Fastl (1990), but in disagreement with an alternative loudness model used by Hellman & Meiselman (1990).

Based on the above simulations and verifications of the model, it was justified, that the model represents known psychoacoustic properties of the normal and impaired human ear, with the exception of temporal properties.

6 OSSQAR: Neural network model.

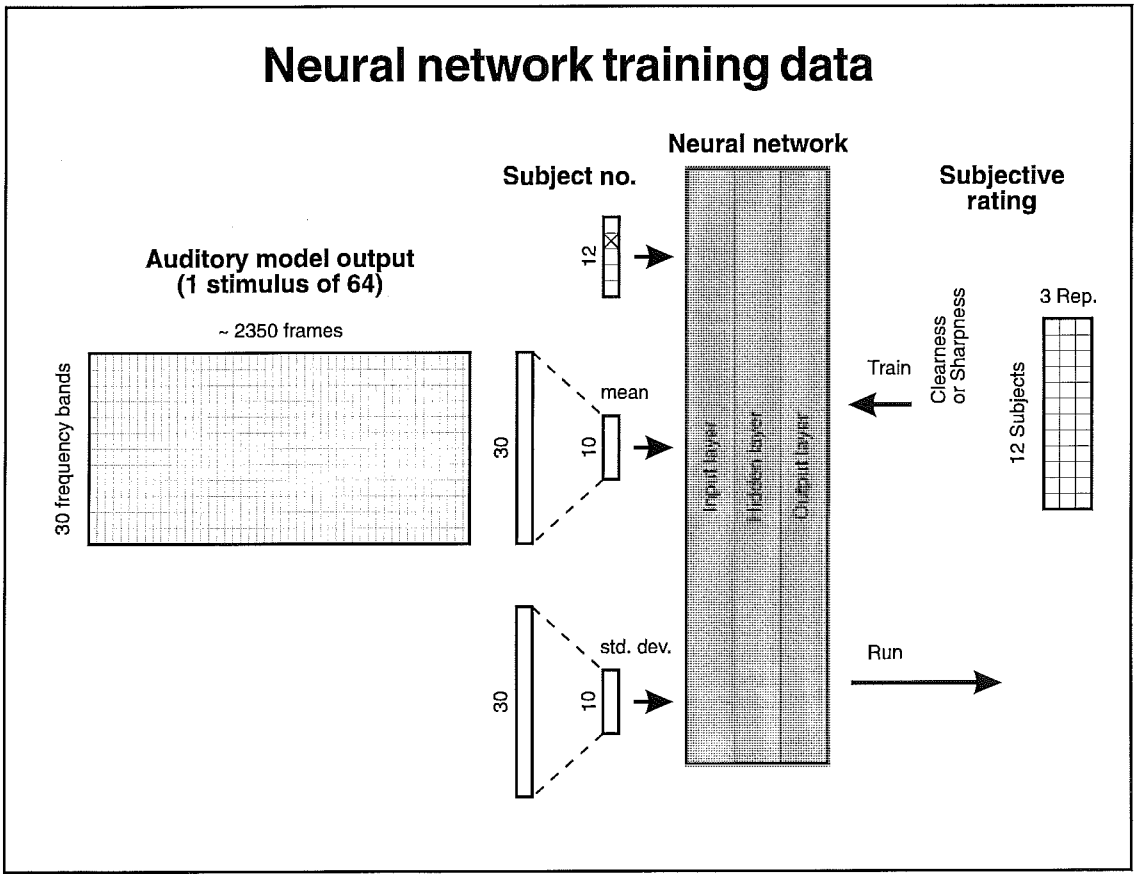
In order to predict the subjective sound quality ratings based on the output from the auditory model, the two sets of data were connected by means of an artificial neural network (ANN). The neural network was then trained using the majority of the subjective rating data and subsequently tested using the remaining subjective rating data for verification. This was documented in detail in Report 3 and summarized below:

6.1 Data representation.

The auditory model processes the incoming signal and outputs the specific loudness (N') in 30 bands, frame-by-frame, every 12.8 ms. Each stimulus file was 30 sec. in duration, but had been presented twice in the subjective rating experiment. By processing the stimuli through the auditory model, 64 output files were generated, each containing roughly 2350 frames, 30 channels wide ≈ 70500 data points, that correspond to one rating of sound quality on each of the perceptual scales. Furthermore, the amount of principally different training data was limited to the number of different stimuli, i.e. 64 different inputs to the auditory model and neural network. Since network size (i.e. the number of free parameters = number of weights) is limited by the amount of training data, the neural network had to be quite small. Hence, the 70500 data points per stimulus had to be reduced to a much smaller number. Ideally this data reduction should not discard any perceptually relevant information from the stimulus, with respect to perceived sound quality.

In the spectral domain, the 30 channels were reduced to 10 by adding adjacent channels 3 by 3. Thus, the spectral resolution which was available to the neural net was reduced. This reduction was justified by analyzing the correlation between output channels for a number of input stimuli. This analysis showed a high amount of correlation between channels close to each other.

For handling of the temporal aspect, various advanced time-dependent neural structures were considered, but a much simpler approach was decided on. For each of the 10 channels from the spectral reduction, the Mean and the Standard Deviation was calculated across all frames, i.e. across the time dimension. The mean values in the 10 bands were assumed to represent the long-term loudness spectrum of the signal and the standard deviation in 10 bands represented the temporal dimension in the 10 bands, or the amount of fluctuations in loudness over the time dimension. The data reduction and representation on both the input and the output side of the neural network is depicted in Figure 15.



15. *Schematic representation of the data inputs and outputs to the neural network and the types of data reduction used to facilitate training. From Report 3.*

Separate networks were trained for the two most prominent rating scales, Clearness and Sharpness (Report 1 and Section 4.3, Figure 13). For each of the 64 stimuli, 12 normal-hearing or 11 hearing-impaired subjects provided 3 ratings each, i.e. 36 or 33

ratings per stimulus. For the neural network these data provide conflicting information because of differences between subjects and because of the large random variability in the subjective ratings. It was assumed that the network could ignore these variations and correctly infer the underlying trend. During the experiments with different network structures and training parameters, 12 input nodes were added to the neural network to identify the subject currently under training. In this way the two largest effect in the subjective rating data, stimulus and subject (Figure 10), were accounted for. The addition of network inputs identifying the subject yielded much better predictions on the test stimuli.

6.2 Neural network implementation.

The neural network was implemented by means of a commercially available software package. The basic structure chosen for the current project was a 3-layer multilayer perceptron, thus a single hidden layer was used. The optimal number of hidden units (neurons) was difficult to determine, but it was set to a number between the input dimension (10 - 32, depending on training session) and the output dimension (1 rating scale). The neural network software had a feature to modify the network structure during training, by adding hidden units to improve the network fitting to the training data. This type of constructive learning was used in all training sessions.

The neural network was trained using the conventional back-propagation algorithm, thus minimizing the mean squared error across the training set. At regular intervals during training, the training was interrupted and the current state of the network was tested by means of the independent test set data. The network state (current weights) was also saved. After many training runs, the optimal state of the network could be determined and the corresponding set of weights could be retrieved.

6.3 Test sets and performance.

The prediction performance of the trained neural networks was determined by means of a test set of data, picked out (removed) from the original subjective rating data before training. Two types of test sets were used in two separate training and test sessions. Usually, test sets are picked randomly from the complete data material. This method was not suited in the present application, due to the 36 (33) replications per stimulus. If some replications from a stimulus ended up in the training set and others in the test set, this would not be a truly separate test set. Therefore, all ratings for selected stimuli were picked out as test set (see Report 1 for a description of the 64 stimuli and Report 3 for documentation of the subset of test stimuli). The test set performance was expected to depend on the way the test set stimuli were picked, so two very different ways of picking them were examined. The first test set of eight stimuli was a **balanced,** or mixed, test set. This way of picking the test stimuli creates a balanced set: 4 with music, 4 with speech, 4 without noise, 4 with noise etc. The second type of test set was picked as all stimuli in **one group**. In the present case, these were all speech signals that were clipped in the mid-frequency band. This was a more biased test set, and consequently the remaining data used for training was also more biased in the opposite direction. The prediction error on the **group test set** was larger, and the sound quality of the test set stimuli was always overpredicted.

Through an evolutional process network size, input data representations and training parameters were changed to optimize the test and training set performances. For the balanced test set, the prediction error was slightly larger than the statistical error in the original subjective rating data. This is documented and discussed further in section 7.

Instead of either of these two test sets, the optimal way to test the model would be with data from a new experiment that used different stimuli and subjects but the same subjective rating scale and procedure. This would also make all of the present rating data available for training.

6.4 Analysis of weights.

It was hypothesized and hoped that the weights in the trained networks could be analyzed to infer the underlying relationship between physical signal parameters and subjective perception of sound quality. The weights to and from the hidden layer were plotted in 3-D contour plots (Report 3) to look for obvious patterns. These plots showed that for certain hidden units, all input weights had been driven into saturation, such that the neuron never responded to any input. This was checked by plotting activation patterns for all units in the network, verifying that the same hidden units were never active ("brain-dead"). Other than this elimination of hidden units by the training, no obvious and meaningful patterns were identified by inspecting the weights. A neural network often has its knowledge distributed throughout the network, thus a single neuron or groups of neurons cannot be linked to particular features in the incoming signal (see also section 8.2).

Another way to view it, is to view the auditory model plus the neural network as a very complex model for prediction of sound quality. A complex model is more likely to make accurate predictions, but by nature, it cannot be expected to provide a simple interpretation. A simple model (i.e. multiple linear regression) on the other hand, might provide a simple interpretation, but poorer predictions.

6.5 Conclusion: Neural network model.

The neural network model implemented as part of the OSSQAR project was successful in predicting subjective ratings of sound quality, when used with an auditory model as pre-processor. The output of the auditory model was reduced in both the frequency and time domains to allow for a reasonably small neural network.

Best prediction results were obtained by providing the neural network with a subject input in addition to input from the auditory model. This allowed for different states in the network for each subject.

The verification of the network was done with a test set picked from the total data set from the subjective rating experiment. The accuracy of the test set predictions depended on how the test set was picked. Using a mix of stimuli for testing showed prediction errors only slightly larger than the random errors in the subjective rating data itself. Poorer predictions were found, using a specific group of stimuli as test set: clipped speech signals. In this case, the neural network tended to overpredict the sound quality on both of the subjective scales: Clearness and Sharpness. A true verification should be performed using data from a new subjective rating experiment with different stimuli and the same rating procedure. Also verification with real hearing aids should be done as an important benchmark, using the types and degrees of distortion and signal processing encountered by the hearing-aid user.

An analysis of the weights in the trained neural networks showed no simple functional patterns that could be used to deduce the qualitative relation between physical parameters in the sound signal and the perceived sound quality.

7 Evaluation of OSSQAR.

This section evaluates the prediction accuracy of OSSQAR and compares it to similar objective methods presented in the review in section 2. The advantages and limitations of applying OSSQAR are discussed.

7.1 Design choices.

With reference to the discussions in sections 1 and 2, certain choices were made in the process of designing OSSQAR:

- Due to the lack of an optimum (transparent) reference system for the hearing aid user, a relative measure was out of the question. It was thus decided to create an absolute measure.
- In order to rate degradations as well as improvements, with no possible separation of the two, a numerical measure was required.
- Perceptual scales (Clearness and Sharpness) were selected rather than "global" scales (e.g. Naturalness, Fidelity), also due to the lack of an optimum reference signal / system, and because they were expected to provide more information.

It was not clear initially, which perceptual scales (or combination hereof) that OSSQAR should predict. In the subjective rating experiment (Report 1), it was found that two principal factors accounted for roughly 70% of the variance. These two factors were by definition orthogonal (not correlated) and thus ideal for training. However, factor I was practically equal to the Clearness scale and factor II was dominated by Sharpness, for both subject groups. Therefore, it was decided to train the neural network on these two rating scales directly, making interpretation of the outcome easier, since the factors I and II had small contributions from the other scales (primarily Fullness). Furthermore, Clearness was highly correlated to Overall Impression, which was then effectively included. The identical interpretation by both subject groups is very important, because

identical objective measures can then be proposed for both groups and be applicable to a range of hearing losses.

As for the statistical analysis of the rating data, no transformation or normalization was made to the rating data before training the neural network. No strong theoretical evidence for such modifications was found, i.e. with the little knowledge we have concerning these perceptual scales, it is best to use the raw data.

7.2 Verification with test set data.

As pointed out during the verification of the neural network performance (Section 6.3 and Report 3), the prediction error on previously unseen data (test set) depends how this test set is picked from the complete set of subjective rating data. In the following, OSSQAR has been evaluated based on the test set with the best performance, the mixed test set. This was considered more representative of a true verification done with data from a new subjective rating experiment with different stimuli. The alternative test set, a particular group of signals, omitted important data from the training set, and the trained networks were thus not as general.

The prediction error on the mixed test set was compared to the error on the stimulus means, given by a 95% confidence interval for the mean from the subjective ratings (Report 1). Few data points exceeded this 95% interval , however the interval used in Report 3 was incorrectly large. That confidence interval was based on the error variance of the stimulus-subject group interaction, which remained when all significant effects and interactions had been removed. Furthermore, it was based on both subject groups together. The final configuration of the neural network had inputs for signal and subject - thus network performance should be compared against the error variance that is left, when these two effects have been taken out. This means that the day effect and related interactions should be pooled with the error variance. And the slightly poorer reliability of the hearing-impaired group should be included.

Hence, the 95% confidence intervals (Tukey Honest Significant Difference) were recalculated for each subject group separately. A similar analysis of variance (ANOVA) had been done previously (Report 1) with the inclusion of a day effect and a subject-day interaction. These two effects were so small, that they did not affect the confidence intervals significantly, which were therefore taken from this 3-way ANOVA (Report 1). The resulting confidence intervals were actually reduced, compared to those used in the neural network evaluation (Report 3).

The correlation between actual and predicted value was characterized by the multiple correlation coefficient R^2 during the neural network experiments (Report 3), which expresses the fraction of variance explained by the multi-dimensional model SS_{mod}, by subtracting the residual variance (SS_{res} = residual Sum of Squares) from the total variance (SS_{tot}) and dividing it by the total variance:

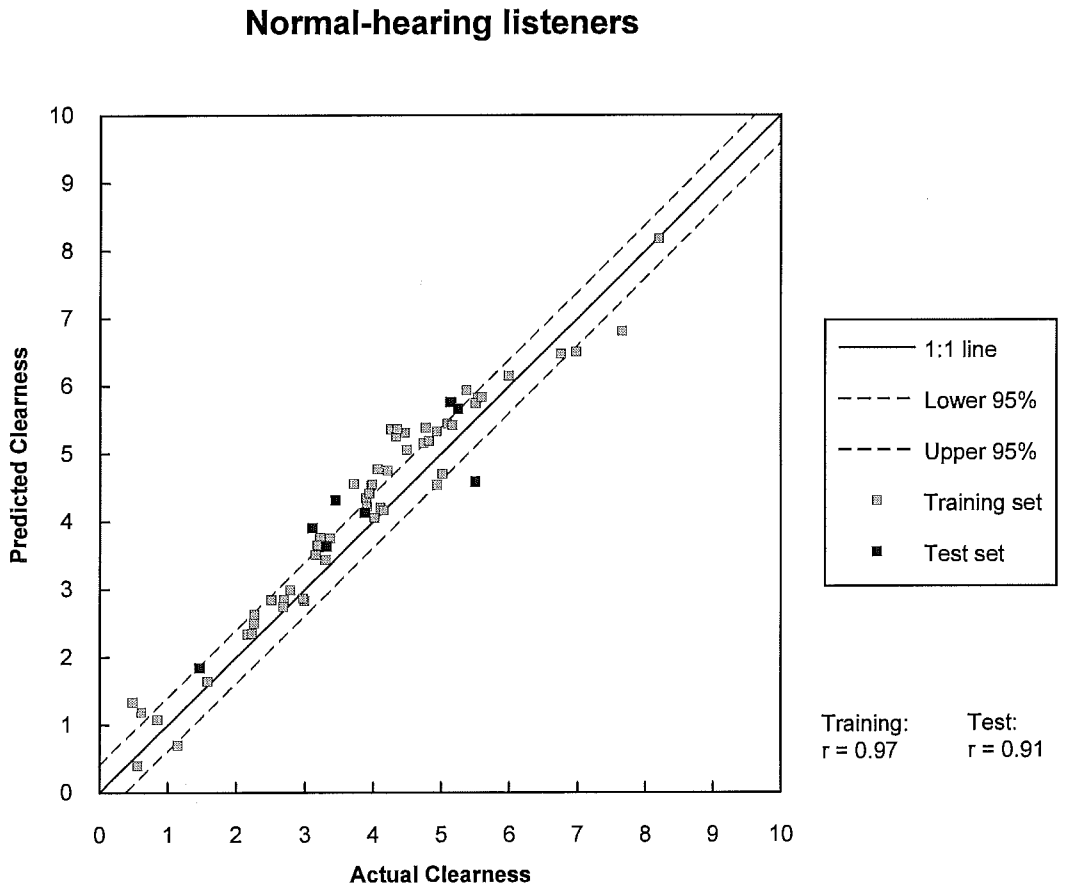
$$R^2 = \frac{SS_{mod}}{SS_{tot}} = \frac{SS_{tot} - SS_{res}}{SS_{tot}} \qquad \textbf{(2)}$$

This measure thus views the actual rating as the *dependent variable* and the predicted rating as the *independent variable*. This has been calculated for both the training sets and the test sets across all subjects. Since the training data itself is noisy, the R^2 will never equal 1 in the present case, but is limited by error variance in the data. The R^2 for the data was thus calculated according to (**2**), using the sums of squares from the 3-way ANOVA (Report 1).

If the actual and predicted values are instead considered *independent*, the prediction performance should be characterized by the correlation coefficient (r) instead. This is the common figure of merit used in the literature. These numbers, on the other hand, tend to be quite large (near 1), if a few points at opposite ends of the scale are provided. The correlation coefficients given for OSSQAR were based on mean predictions and mean actual ratings taken across all subjects in the group.

7.3 Prediction performance.

The prediction performance was evaluated for Clearness and Sharpness, and for the two subject groups separately. Only mean values of predicted and actual values are shown here. Scatter plots of all predictions are shown in (Report 3). The plot of actual vs. predicted values of Clearness for the Normal-hearing group is shown in Figure 16.



16. OSSQAR prediction of Clearness vs. the mean actual rating for the 12 normal-hearing subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.

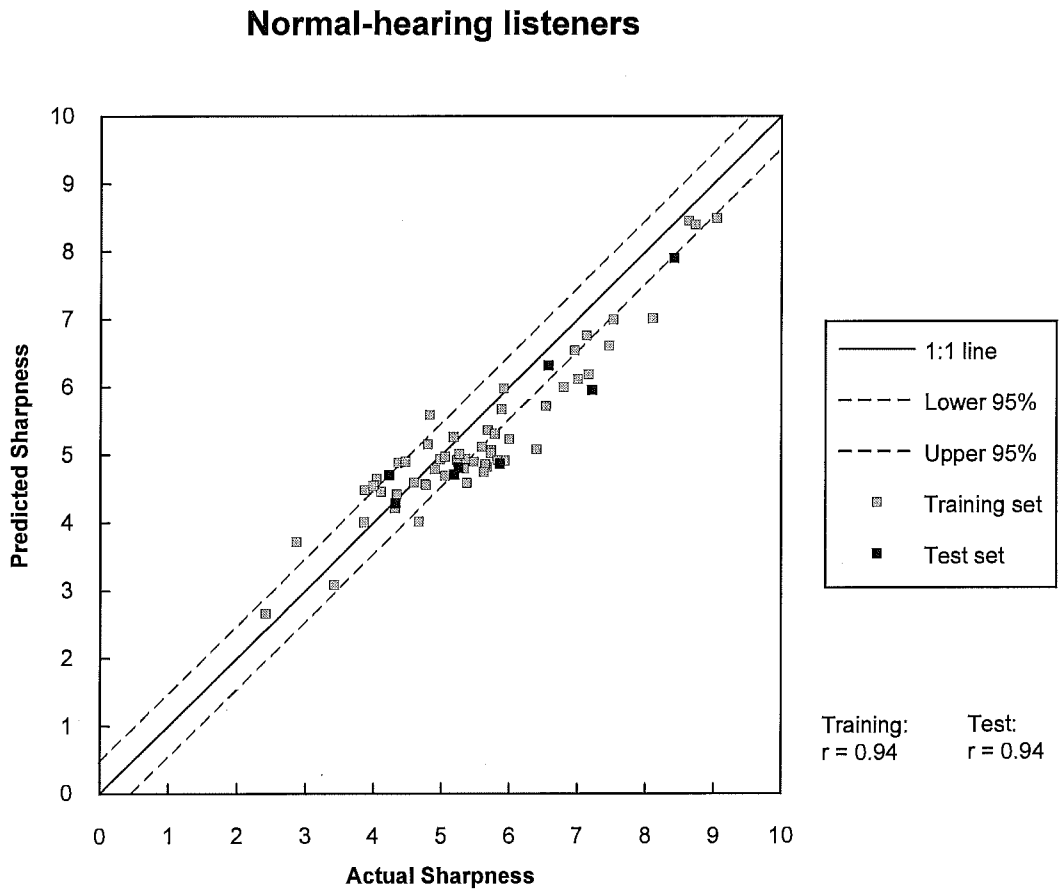
In this and the following plots, the predicted value was considered the dependent variable, thus it was plotted along the y-axis. The actual rating was considered the

reference, or the independent variable, and plotted along the x-axis. The 95% confidence intervals could have been plotted as horizontal error bars from each point, but this gave a very cluttered plot. Instead, the confidence intervals were plotted as dashed lines with that same horizontal deviation from the 1:1 line that represented perfect prediction. Data points falling outside these lines thus have a significant prediction error.

For Clearness shown in Figure 16, the predicted values of the training data are scattered in a symmetrical band around the 1:1 line, with some points outside the confidence intervals. The same picture is seen for the test set, with about the same amount of prediction error. The largest test set error is -0.9 (data point below the 95% line), for stimulus 55, which is speech with background noise, clipped in the low- and high-frequency bands and unchanged in the mid-frequency band (Report 1). Generally, there is a little overprediction of Clearness in the middle of the scale.

The correlation is high ($r = 0.95$) for the training set and slightly lower for the test sets ($r = 0.92$). These are similar to the values provided by Herre et al (1992), who predicted subjective degradation on a 1-5 scale. Their prediction values should be compared to the above plot of Clearness, since Clearness is almost identical with Overall Impression (Report 1). The *training set* prediction values from Herre et al (1992) are shown in Figure 5. No verification with independent test data was done. Their maximum prediction error is 0.5 (12.5% of full scale), which can be compared to the present maximum training set error of 11% and maximum test set error of 9% for Clearness, as shown in Figure 16.

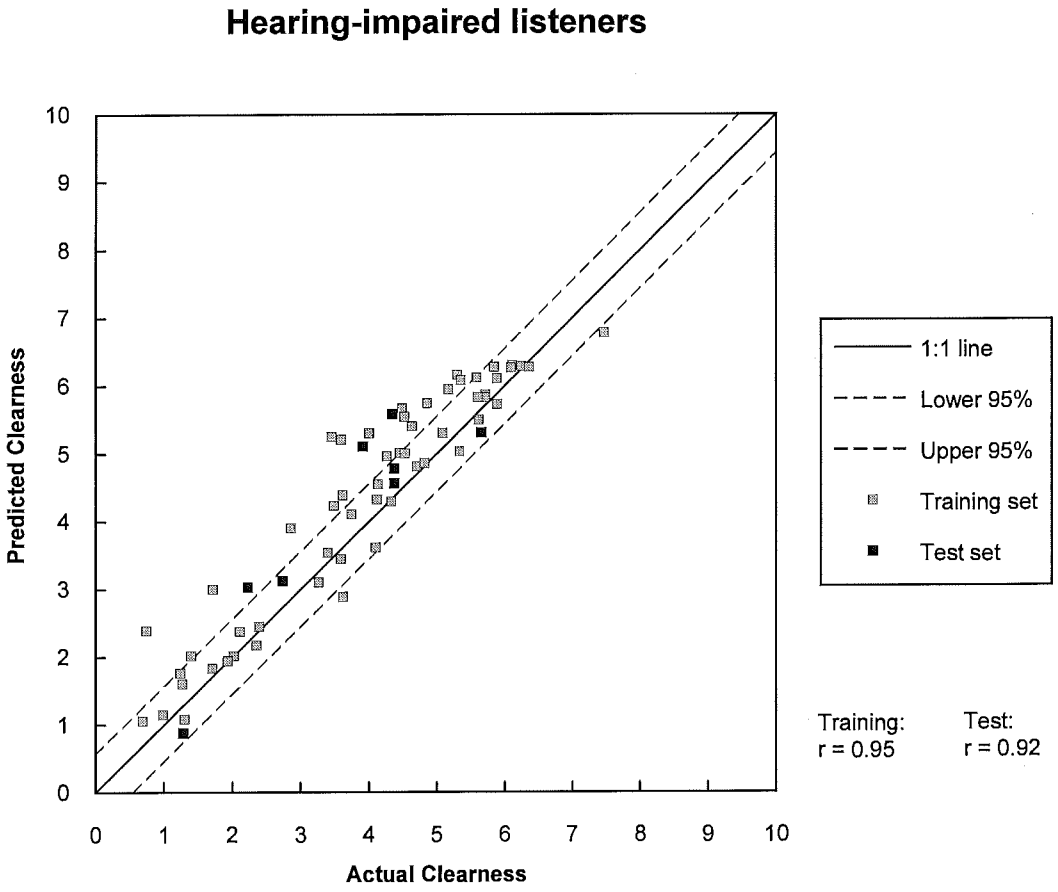
The predicted values of Sharpness, by Normal-Hearing listeners, is shown in Figure 17.



17. OSSQAR prediction of Sharpness vs. the mean actual rating for the 12 normal-hearing subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.

There is some underprediction of Sharpness in the range 5 - 9, i.e. the "poor" side of the scale. The training data are not evenly spaced along the Sharpness scale, and thus not optimal for training. The test and training set errors are very similar, and the two correlation coefficients are identical (r = 0.94). The maximum deviations are 13% on both training and test sets.

No similar published results are available for comparison. Von Bismarck (1974a) has published an objective measure of Sharpness, but no comparison between actual and predicted values has been done.

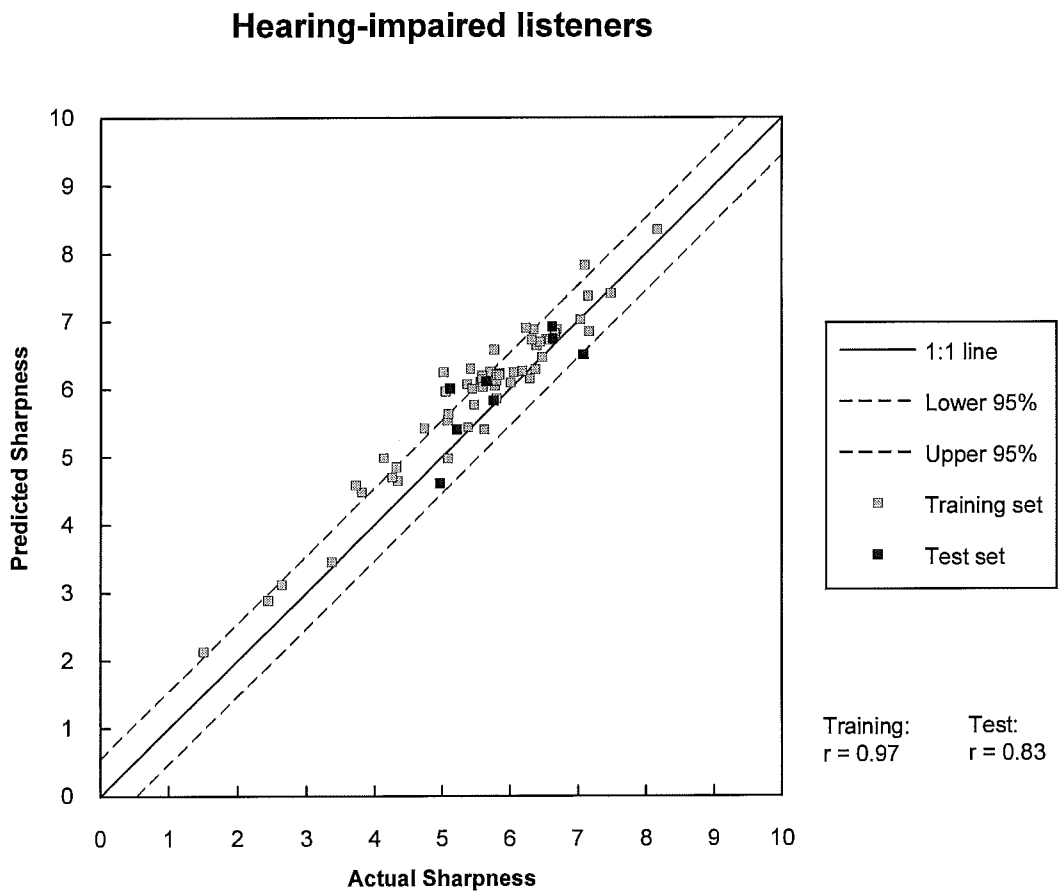


18. *OSSQAR prediction of Clearness vs. the mean actual rating for the 11 hearing-impaired subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.*

The predicted values of Clearness, by Hearing-Impaired listeners, is shown in Figure 18. The spread around the 1:1 line is larger than for the NH group (Figure 16), but so is the 95% confidence interval, and roughly the same number of stimuli fall outside of the 95% limits in the two cases (26 for NH, 21 for HI). The correlation coefficients are good, $r = 0.95$ for the training set and $r = 0.92$ for the test set. The maximum prediction

errors are larger that for the Normal-Hearing subject group: 18% for the test set and 13% for the test set. The test set is generally predicted with the same accuracy as the training set.

The predicted values of Sharpness, by Hearing-Impaired listeners, are shown in Figure 19.



19. *OSSQAR prediction of Sharpness vs. the mean actual rating for the 11 hearing-impaired subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.*

As for the NH group, the actual Sharpness ratings are clustered close to 5, and the training data are thus not perfect. There is generally a small overprediction of

Sharpness, unlike the underprediction in the NH case (Figure 17). However, the spread outside of the 95% limits is smaller than for the NH subjects, only 16 points are outside the limit, compared to 31 for the NH group. This is also reflected in the larger correlation coefficient for the training set ($r = 0.97$). The test set correlation is moderate ($r = 0.83$), due to a very clustered test set.

The OSSQAR prediction performance has been summarized in the table in Figure 20, along with similar performance measures for other objective quality measures.

Objective measures of sound quality.

	r Fitted data	*r* Test data	R^2 Training	R^2 Test	R^2 ANOVA
OSSQAR: NH Clearness	0.94	0.94	0.73	0.49	0.68
OSSQAR: NH Sharpness	0.97	0.91	0.54	0.48	0.51
OSSQAR: HI Clearness	0.95	0.92	0.58	0.4	0.53
OSSQAR: HI Sharpness	0.97	0.83	0.6	0.53	0.47
NMR Herre et al, 1992	0.94				
PAQM Beerends & Stemerdink, 1992	0.97	0.91			
PERCEVAL *) Paillard et al, 1992	0.92				
Composite Acceptability Quackenbush et al, 1988	0.84				

20. *Table of performance measures for a number of objective quality measures. The correlation coefficient between predicted and actual values, r, is given for fitted data (used for the training / design of the measure) and for test (verification) data, not used during model building. R^2 values have not been presented in the literature and are provided to compare OSSQAR performance between subject groups and scales. *) PERCEVAL correlation was taken from Beerends & Stemerdink, 1992.*

This table should be interpreted with some caution. The types of distortion and divergence of test stimuli are very different across the four studies. However, PAQM and PERCEVAL were evaluated on the same database with subjective ratings of bit-rate reduction codecs. The poorer performance of Composite Acceptability (Quackenbush et al, 1988) is probably due to a larger and more diverse set of distortion types. OSSQAR, on the other hand, was evaluated on a test set very similar to the training set, but

independent. The only other independent verification was done for PAQM, with some degradation of correlation from training to test (Beerends & Stemerdink, 1992). The correlation coefficients are generally on the same order, indicating similar performance, but these figures should not be scrutinized further, and they do not suffice as a thorough evaluation of the objective measures. See the following section for further discussion of the validity and applicability of OSSQAR and other objective measures.

With the evaluations possible for OSSQAR, we also have the multiple correlation coefficient, R^2, as an indicator of prediction performance. The optimal value, $R^2 = 1$, can only be reached if the desired outputs in the training data are free of random noise. The theoretical R^2 was calculated from the three-way ANOVA, by dividing the sums-of-squares due to stimulus and subject with the overall sum-of-squares (**2**). The complete set of subjective rating data was included, and the result is shown in the rightmost column of Figure 20. For the neural network training and testing, the complete data set was split into the training and test sets. The corresponding values calculated by the neural network software are shown in two columns in Figure 20. The values for the training set are slightly above the theoretical value, indicating as good training as possible with the available data. The unexpectedly high R^2 may be due to a selection of a favorable data set as training set. The corresponding values for the test set are below the ANOVA values, indicating less than optimal generalization performance. An exception for this is the test performance for Sharpness by the hearing-impaired group. It is not clear why both training and test set performance in this case exceeds the R^2 based on ANOVA.

7.4 Discussion: OSSQAR

As documented in Figures 16 - 19, it is indeed possible to successfully develop and train a model for prediction of sound quality. The prediction results are very good and indicate that further work could be worthwhile. A number of issues concerning

OSSQAR deserve further attention, relative to the measures discussed in the literature section (section 2) and the design goals set for OSSQAR in section 3.

Absolute measure. OSSQAR was designed to be an absolute measure of sound quality, independent of any optimum or given reference. However, the basis of OSSQAR are the subjective rating scales, so the main question is, if these are absolute. Analysis of the subjective data (Report 1: Table IV) showed that the overall mean ratings on each scale were equal for the two subject groups. It is reasonable to assume that the auditory impression is not the same for the two groups, because the sensorineural hearing loss was only compensated for by linear amplification. And this difference was not reflected in the subjective ratings, on any of the scales, which must then be considered relative to the overall experiment mean. We can hope that the context dependence has been reduced by designing an experiment with very diverse stimuli. It is likely that OSSQAR is about as absolute as a sound quality measure can be.

Ideal values. No explicit ratings of ideal values were obtained during the subjective listening tests, and such hypothetical ratings should probably be interpreted carefully. Instead, we will examine which stimuli obtained the best ratings of Overall Impression. These stimuli were signals processed as little as possible, i.e. no or little degradation was done to these signals. For both subject groups, this condition was for stimulus number 61: speech, with no noise, no filtering and no clipping or compression, i.e. completely clean speech. The means on the Clearness, Sharpness and Overall Impression scales are shown in Figure 21.

Stimulus no.	NH Clearness	NH Sharpness	NH Overall	HI Clearness	HI Sharpness	HI Overall
61	8.2	4	8.5	7.5	5.1	7.5

21. Mean ratings of the stimulus with best rating of Overall Impression, which was the same for both subject groups. Compare to the rating scales and the verbal fixpoints shown in Figure 9.

Given the verbal fixpoints on the Overall Impression scale (9: Very good), it is safe to assume this end as an optimum, i.e. 10 represents the best quality. The clearness scale may have an optimum at a lower point, i.e. at 10 the reproduction sounds "too Clear". Stimulus 61 received a Clearness rating of 8.2 and 7.5 by the two groups, respectively. The likely optimum on this scale is thus in the range 7 - 9, which was also found by Gabrielsson and Hagerman (1993).

The Sharpness ratings for this stimulus are probably also close to optimal sound quality. The NH group rated Sharpness at 4.0 and the HI group rated Sharpness at 5.1. Most likely, the optimum on this scale is in the range 4 - 5. Gabrielsson and Hagerman (1993) found 5 or slightly above as the optimum for Softness/Gentleness, which has the inverse direction compared to Sharpness. If mirrored around 5, the midway point, the present ideal values are the same.

Given the concerns about the absoluteness of OSSQAR and the uncertainty about the precise location of ideal values (optimum) on both the Clearness and Sharpness scales, it is difficult to provide exact rules on how to use these measures. If the objective estimates are far from the optimal values mentioned above, there is most likely a serious problem with the sound quality in the device under investigation. In such a situation, OSSQAR can be used to rank a number of conditions relative to the optimum point. Closer to optimum, it becomes difficult to use OSSQAR for refinement of the sound quality. These types of problems are the same as for traditional subjective evaluations, and thus not a specific weakness for the objective measures.

Individual variations. On both the Clearness and Sharpness scales, for both subject groups, there were significant differences between mean ratings of individual subjects (Section 4.3, Figure 10 and Report 1). This means that some subjects generally rate higher on a scale than other subjects, either due to different interpretation of the scales or different preferences. The predictions made by OSSQAR are averages for a population, and thus only a guideline for predicting the perceived sound quality of an individual, i.e. for fitting of hearing aids. This is underlined by the limitations

mentioned above. However, for development of signal processing concepts for the hearing impaired and hearing aids, we want to make statements about the general population. OSSQAR is well suited for this.

Future applications of OSSQAR. Given the above considerations, there are limitations to the use of OSSQAR. First, and foremost, new verification experiments are required to verify the validity of the objective measure. Even then, OSSQAR will never become more correct or absolute than its subjective basis. It is unlikely, that OSSQAR or other objective measures will replace subjective ratings of sound quality. They should be viewed as a supplement to the subjective methods. As such, they can be used to guide the scientist or engineer in the right direction and for instance help design the proper subjective tests, by sorting out obvious differences, and selecting only critical stimuli, whereby valuable experimental time can be saved.

OSSQAR also represents a substantial improvement over the traditional physical "quality" measures, such as frequency response, signal-noise-ratio, distortion, but it should not be viewed as a replacement for these methods neither. It is a supplement to the traditional objective measures.

The informal listening test is another common development tool that should not be neglected. Whether OSSQAR is more or less sensitive than such tests depends on many factors, which could perhaps be examined in a future experiment. However, designing a scientifically sound experiment to examine the potential of informal listening is a contradictory task. OSSQAR has the advantage over informal listening, that the bias of the subject is removed, and that hearing loss is included correctly. The auditory model itself, without the neural network model, will also be a valuable tool in the hearing aid development process, due to the inclusion of hearing loss.

8 From the physical to the subjective domain.

One of the purposes of the present project was to explore the relations between, on one side, the physical parameters of the signal and the reproduction, and on the other side, the subjective sound quality, perceived by the hearing-aid user. Two results facilitate this: **1**) The qualitative results from the subjective listening tests, where signal and processing parameters were combined in a systematic way, using a factorial experiment (Report 1) and **2**) the weights in the trained neural networks that performed the mapping from a psychoacoustic representation of the signal, provided by the auditory model, to the sound quality, perceived by the subjects.

8.1 Signal and processing effects (factorial analysis).

The 64 stimuli were designed as a systematic combination of different input signals and parameters (Report 1) as shown in Figure 22 below:

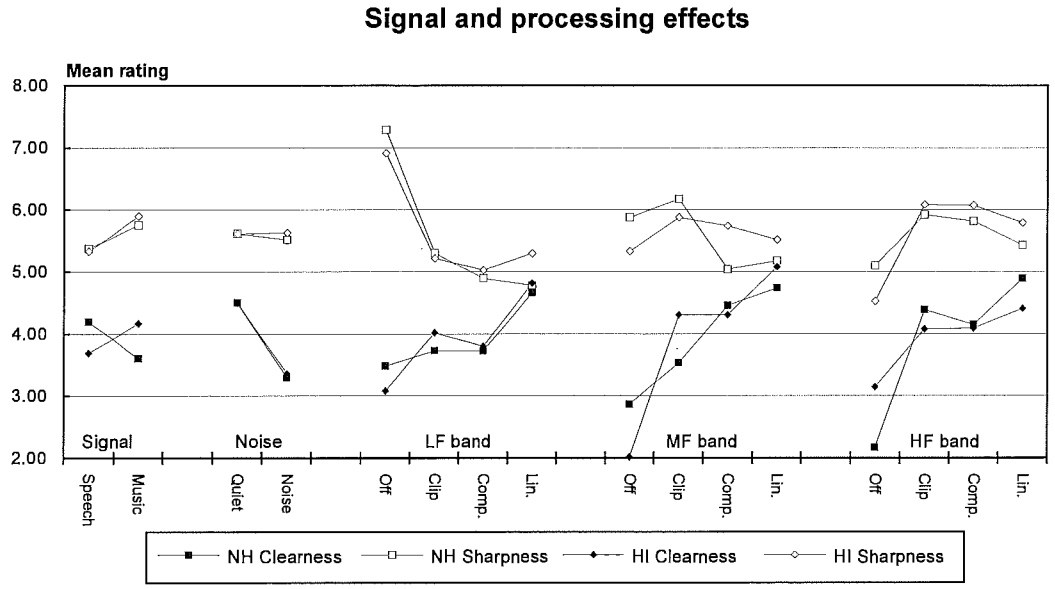
	Input signal	Background noise	Low-frequency band 0 - 500 Hz	Mid-frequency band 500 Hz - 4 kHz	High-frequency band 4 kHz - 10 kHz
Level 1	Speech	Off	Off	Off	Off
Level 2	Music	On	Clip	Clip	Clip
Level 3			Compress	Compress	Compress
Level 4			Linear	Linear	Linear

22. Listing of input signals and processing conditions used in factorial experiment. See section 4.1 and Report 1 for details on the stimuli.

The input signal is either speech or symphony music, mixed with background noise or not, and then split into three frequency bands. Each band is either switched off, clipped, compressed, or fed straight through (Linear). The stimulus is then formed by summing the outputs of the three bands.

In a completely crossed experiment using the above options, a total of 2*2*4*4*4 = 256 stimuli are produced. However, a reduced number of 64 stimuli have been picked to form a so-called fractional factorial experiment (Box et al, 1978). This experiment still

allows us to estimate the effects of changing the signals and processing parameters on the rating scales (main effects), only higher-order interactions are confounded. The detailed effects and tests for significance are listed in (Report 1). In the following we will summarize qualitatively the effects on the two prominent rating scales used in OSSQAR. The mean ratings for all signal and processing parameters are shown for both subject groups in Figure 23.



23. Plot of means for all signals and processing parameters, used in the subjective rating experiment. Both normal-hearing (NH) and hearing-impaired (HI) subjects are included. The range on the vertical axis has been reduced from the 0-10 range to provide a better view of the mean values. Adopted from Report 1. See text for details.

The following observations can be made:

Input signal: The normal-hearing group rate music lower than speech in terms of *Clearness*. The opposite is the case for the hearing-impaired group, but the effects in both cases are small (+/- 0.5). *Sharpness* is increased for music, which is probably due to more high-frequency energy in this signal, in particular in string passages of the symphonic music.

Background noise: The same picture for both groups - *Sharpness* is unchanged, while *Clearness* is severely reduced when background noise is added - this effect should be predictable with OSSQAR, since the temporal fluctuations in each bands are diminished, when background noise is present. The subjects were asked to rate the reproduction only - this is obviously not what happened, since the noise effect was present. Just as the auditory model and the neural network, the subjects are not able to separate input signal from reproduction. This shows a good correspondence between the subjective and objective measures.

Low-Frequency band (0 - 500 Hz): *Clearness* is highest when this band is unmodified (Linear), and reduced for the other processing conditions, thus temporal changes to the signal affect the perceived Clearness. The HI group is more sensitive to removal of this band (Off), probably because they have the least hearing loss on this range, and thus rely more on it. *Sharpness* is drastically increased when the LF band is turned off, and unaffected by the other changes, thus it is mostly affected by the spectral change. This effect is the same for the two subject groups.

Mid-Frequency band (500 Hz - 4 kHz): There is a general degradation of *Clearness* from Linear to Compressed to Clip to Off. Both spectral and temporal changes are thus detected. For the hearing-impaired group, there is no difference between Clipping and Compression. A possible explanation for this is the high-frequency hearing loss making this group insensitive to the high-frequency harmonics generated by clipping. When the MF band is removed, Clearness is very low for the HI group. They must be expected to rely more on this frequency band. *Sharpness* is not dramatically affected for the HI group, while the NH group rates it higher, when the signal is switched Off or clipped. This is probably due primarily to the spectral changes in these two conditions - removal of energy or adding of high-frequency harmonics.

High-Frequency band (4 kHz - 10 kHz): *Clearness* is reduced dramatically, when this band is removed (Off), in particular for the NH group. For both groups, the Linear condition was superior to the clipped and compressed conditions by a small margin.

Thus, Clearness is mostly a spectral effect in this frequency band. *Sharpness* is reduced when this band is turned off, and otherwise affected only slightly, so this must also be considered a spectral effect.

Generally, there were little differences between the two subject groups, confirming that the meaning of the scales and the perception of sound quality are similar for the two groups. The only major difference is a higher sensitivity for the Normal-Hearing group.

The above observations indicate that *Clearness* is affected by both spectral and temporal properties. High *Clearness* requires full bandwidth, as Gabrielsson and Hagerman (1993) also pointed out.

Sharpness is mostly affected by spectral changes in the present experiment, both in the high and low frequencies - similar to the results of von Bismarck (1974b), that based Sharpness on the center of gravity in the specific loudness pattern. Gabrielsson and Hagerman (1993) found that *Sharpness* was only affected by high-frequency changes.

8.2 Auditory model and neural network interpretation.

The auditory model and the neural network together form a complicated calculation model for prediction of sound quality, OSSQAR. By analyzing this model, after training, it should be possible to provide a more detailed picture of the physical-subjective relation, than from the factorial experiment discussed above.

A quantitative interpretation of the network would be a valuable result, but such an interpretation might be in contradiction with the concept of a complicated model. The network weights are quantitative measures, but there are many of these and they interact through the nonlinear processing elements in the network. A linear regression model, for example, has a simple quantitative interpretation, but this model may not be able to fit the data well. Thus, a quantitative interpretation is not possible nor meaningful based on neural nets.

The network weights were plotted in 3-D contour plots to be used for a qualitative interpretation of the trained model, by interpreting the visual patterns (Report 3). These patterns showed no clear trends, like for instance neurons that were sensitive to the slope of the Spectrum for prediction of Sharpness. It was concluded that the neural network was a distributed model, that provided good predictions, but is impossible to interpret in simple terms. This outcome is not uncommon in neural networks: We have successfully developed a "black box" for prediction of sound quality, and it remains a black box!

Instead, we can look at the input-output relation of the network for certain stimuli. This was done for a few stimuli, that represented the extremes on the Clearness and Sharpness scales (Report 1 - Table X). These patterns provided no new knowledge concerning the physical-subjective relationship, in addition to the qualitative results in section 8.1.

Further examination of the input - output relation of the neural network may be worthwhile in the future. For instance by analyzing the input and output patterns for the present stimuli. Or by feeding the trained network with hypothetical inputs representing different extremes of input signals, i.e. spectra with different slopes and levels and different amounts of temporal fluctuations. There is one pitfall with this approach: By simulating auditory model outputs in this way, it is possible to generate network inputs that have no physical meaning, i.e. a signal with no average energy (Mean) in a band and large amounts of temporal fluctuations (Std. Dev.) in the same band. This should be avoided.

9 Discussion.

The present project has shown that an objective measure of sound quality is a feasible concept. Such a measure (OSSQAR: Objective Scaling of Sound Quality And Reproduction) has been proposed, implemented and evaluated with good results. Three main elements were addressed in the project:

Subjective evaluation of sound quality by normal-hearing and hearing-impaired listeners (Report 1). The rating experiment provided data covering a large range on the two most dominant perceptual scales : Clearness and Sharpness. These two scales were equally important for the two subject groups, Normal-hearing and Hearing-impaired, and were interpreted the same way. This was an important basis for the development of an objective sound quality measure common to both normal-hearing and hearing-impaired subjects.

An auditory model with hearing loss (Report 2). This model was developed as a practical, quantitative model with a reasonably low processing time. Simulation results from the auditory model were compared to psychoacoustic data from the literature, and there was reasonable agreement. No temporal effects (temporal integration, postmasking) are presently included in the model, and the current application, OSSQAR, indicated that this was not crucial for prediction of sound quality. In fact, most temporal information was discarded by the proposed data reduction (Section 6.1 and Report 3). The primary purpose of the model was to serve as a preprocessor for the neural network, but it was designed to be used separately as well, for the study of perception of complex and real-world sounds.

A neural network model for the prediction of sound quality (Report 3). The high amount of data from the auditory model was reduced to match a smaller neural network. After training with a part of the subjective quality ratings, the network was able to predict the perceived sound quality of the remaining part of the data (test set) with an error slightly larger than the statistical error in the subjective rating data itself. The

prediction error was dependent on how the test set was chosen, and further verification with completely new rating data should be done in the future. The trained networks were examined to discover the underlying relation between signal parameters and subjective quality, but this analysis showed that the inferred knowledge was distributed throughout the network. No simple, qualitative interpretation of the model was thus possible.

The resulting objective measure, OSSQAR, provided a prediction performance similar to other objective sound quality measures, with correlation coefficients on the order of 0.95, but this should not be taken as an indication of a perfect objective measure. More verification is needed on a diverse base of signals and reproductions, to assess precisely the limitations of OSSQAR.

Some limitations were found in the present work: OSSQAR was designed to be an absolute measure of sound quality, but it is no more absolute or simple to interpret than the subjective ratings on which it is based. OSSQAR is only absolute, in the sense that no reference signal is required to predict the sound quality. Furthermore, there was no well-defined optimum on the two perceptual scales that are predicted by OSSQAR: Clearness and Sharpness. Therefore, it can only be used to guide in the right direction, when developing hearing aids, but not to point out, when the optimum reproduction quality has been reached.

OSSQAR makes predictions for a population of listeners, and since there are significant inter-individual differences in ratings and preferences (Figure 10), it should generally not be applied to hearing aid fitting for an individual.

Given these limitations, OSSQAR cannot replace any of the traditional methods: subjective listening tests, electroacoustic measurements, informal listening tests, but it has a place as a supplement to these. It can help design better hearing aids by providing quick results and help guide a researcher to design the optimal subjective test, and avoid wasting valuable test subject time.

The structure of OSSQAR provides another advantage compared to subjective assessment - it is possible to analyze the various stages in the model (auditory model, neural network) for a given signal and to explore what the underlying mechanisms and problems are. With traditional subjective ratings, the result is basically a number without any explanations of the underlying reasons.

10 Conclusion.

A method for the objective estimation of sound quality has been developed and evaluated. It consisted of three components: Subjective sound quality ratings to provide reference data, an auditory model with hearing loss, coupled to an artificial neural network, that was trained to predict the sound quality ratings. The present work has shown that such a measure is a feasible and meaningful concept for both normal-hearing and hearing-impaired listeners, providing fast and repeatable estimates of sound quality.

This measure, OSSQAR (Objective Scaling of Sound Quality And Reproduction), predicts the perceived sound quality on two independent perceptual rating scales: Clearness and Sharpness. These two scales were shown to be the most relevant for assessment of sound quality, and they were shown to have the same perceptual meaning for both normal-hearing and hearing-impaired listeners.

Using test data from the subjective rating experiment, the prediction error of OSSQAR was found to be only slightly larger than the random variance in the subjective ratings. Analysis of the neural network after training did not provide qualitative knowledge about the relation between physical signal parameters and perceived sound quality.

OSSQAR was designed as an absolute measure, however the subjective sound quality ratings on which it was based, were found not to be absolute. Thus, the OSSQAR predictions can be used to rank the quality of the reproductions of hearing aids, but not to predict precisely the outcome of any subjective quality rating experiment.

Further verification with new signals and distortion types will be required to assess how general and reliable OSSQAR is, and to identify the precise limitations of its application.

11 Suggestions for future work.

In the course of a very broad project, like this one, many ideas come up that cannot be pursued further due to time limitations, or because they are outside of the scope of the project.

First and foremost, any objective measure, like OSSQAR, must be validated thoroughly, before we can use it with full confidence. As already pointed out, OSSQAR should be verified with new, independent subjective rating data. These data must of course use the same perceptual scales and rating procedure as used in the present rating experiment. All other experimental parameters should be varied, i.e. different input signals, different types of distortion and different subjects with different hearing losses. A logical test and benchmark would be to compare OSSQAR ratings with subjects' ratings of real hearing aids, and verify if the ratings are the same, and test, which measure is the more sensitive. In this case, OSSQAR could prove to be more sensitive, since it was trained on more diverse stimuli, as used in the original rating experiment (Report 1).

With respect to the subjective evaluation of sound quality, there are many problems still outstanding. Measures that are as free of context as possible, coupled with well-defined ideal values on the scales, are important both for subjective and objective evaluations and would ideally allow for comparisons between experiments. The corresponding objective sound quality measures would thus also become more general.

Given the difficulty of finding truly absolute quality measures, other options could be examined. Although ruled out for hearing aid applications at the outset of the present project (section 3), a relative, objective measure could perhaps be designed as a counterpart to paired comparison with preference judgments or similarity ratings (see Figures 1 and 2). Preference judgments have been used successfully for hearing aid evaluation (Punch et al, 1980), and the task of searching for optimal quality works well in a type of paired comparison tournament (Levitt et al, 1978b). Some type of auditory distance measure that predicts preference could perhaps be developed. Such a relative

quality measure is makes it difficult to compare reproductions, due to the lack of an obvious reference for hearing-impaired subjects. As a development tool, such a measure would be more cumbersome to apply, since the developer would have to use the model as test subject in a paired comparison experiment, and hope that the best possible result was obtained when no other challenging algorithms/hearing aids were preferred.

A major conceptual problem encountered in the present project was how to handle the fluctuating characteristics of the stimulus and match it to the corresponding single rating of sound quality. Just as an auditory model looks at the instantaneous stimulus, the perception of Sharpness will fluctuate over time, as the spectrum of the stimulus changes. Thus, the subjective rating of quality could also be performed as a function of time, for instance by letting the subject move a slider for Sharpness, while listening to the stimulus (Hagerman & Gabrielsson, 1991).

The auditory model section of the present project deserves further attention as well. The assumption that an auditory model is advantageous for prediction of quality is dominant in the literature. This has not been tested by, for instance, replacing the auditory model by a simple preprocessor, like a 1/3-octave filterbank or an FFT, and training the neural network with these inputs. The present auditory model could be enhanced with known psychoacoustic properties, such as temporal integration, post-masking, binaural release from masking (in the case of the more common binaural listening situation).

While working with the auditory model, the author became aware of how limited our knowledge in psychoacoustics is, when real signals are considered. Most of the present psychoacoustic results and models are based on simple signals, and these must then be extrapolated and integrated to form a complete auditory model. Hopefully, auditory models will continue to improve, as the knowledge of perception of sound grows.

The present version of the auditory model predicts the auditory filter shape from the hearing loss, although there are large variations in this and other psychoacoustic

parameters depending on hearing loss. Either these parameters should be measured individually, or at least the range should be known. How much they affect the perceived quality is also not clear, and they may not be critical. The answer to these questions also requires further psychoacoustic research.

The neural network model could be developed further, by experimenting with different data representations to the NN and different network structures, including temporally dynamic networks.

12 References.

Allen, J.B. (1985). Cochlear modeling. IEEE ASSP Magazine, January 1985.

Bech, S. (1987). Listening Tests on Loudspeakers. Report no. 43, The Acoustics Laboratory, Technical University of Denmark.

Beerends, J.G. and Stemerdink, J.A. (1992). A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation. J Audio Eng. Soc., 40 (12), 963 - 978.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics for experimenters. An Introduction to Design, Data Analysis and Model Building. Wiley-Interscience, New York.

Brandenburg K. (1993). Perceptual Models for the Prediction of Sound Quality of Low Bit-Rate Codecs. Proc. AES 12th conference, "Perception of Reproduced Sound", Copenhagen, Denmark.

Brandenburg K. and Sporer, T. (1992). "NMR" and "Masking Flag": Evaluation of Quality Using Perceptual Criteria. Proc. AES 11th conference, Portland, Oregon.

Cohen, J (1989). Application of an auditory model to speech recognition. J Acoust Soc Am, 85(6), 2623 - 2629.

Dubno, J.R. & Schafer, A.B. (1992). Comparison of frequency selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners. J Acoust Soc Am, 91(4), 2110 - 2121.

Dyrlund, O. (1992). Coherence measurements in hearing instruments. Using different broad-band signals. Scand Audiol, 21(2), 73 - 78.

Ferguson, G.A. and Takane, Y. (1989). Statistical Analysis in Psychology and Education. McGraw-Hill, New York.

Gabrielsson, A. and Hagerman, B. (1993). Subjective correlates of the acoustical characteristics of sound-reproducing systems. In: Acoustical factors affecting hearing aid performance (eds. Studebaker, G.A. and Hochberg, I.). Allyn and Bacon, Needham Heights, MA.

Gabrielsson, A., Schenkman B.N. & Hagerman, B. (1988). The effects of different frequency responses on sound quality judgments and speech intelligibility. Journ Speech Hear Res. 31, 166 - 177.

Gabrielsson, A and Sjögren, H. (1979). Perceived sound quality of hearing aids. Scand Audiol, 8, 159 - 169.

Glasberg, B.R. and Moore, B.C.J. (1986). Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. J Acoust Soc Am, 79(4), 1020 - 1033.

Glasberg, B.R. and Moore, B.C.J. (1990). Derivation of auditory filter shapes from notched-noise data. Hear Res 47, 103 - 138.

Grewin, C. (1993). Methods for Quality Assessment of Low Bit-Rate Codecs. Proc. AES 12th conference, "Perception of Reproduced Sound", Copenhagen, Denmark.

Hagerman, B. and Gabrielsson, A. (1991). Personal communication.

Hellman, R.P. and Meiselman, C.H. (1990). Loudness relations for individuals and groups in normal and impaired hearing. J Acoust Soc Am, 88(6), 2596 - 2606.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am, 87(4), 1738 - 1752.

Herre, J., Eberlein, E., Schott, H. and Schmidmer, C. (1992). Analysis Tool for Realtime Measurements Using Perceptual Criteria. Proc. AES 11th conference, Portland, Oregon.

Kapust, R. (1992). A Human Ear Related Objective Measurement Technique Yields Audible Error and Error Margin. Proc. AES 11th conference, Portland, Oregon.

Karjalainen, M. (1984). Sound quality measurements of audio systems based on models of auditory perception. Proc. ICASSP 1984, San Diego, California.

Karjalainen, M. (1985). A new auditory model for the evaluation of sound quality of audio systems. Proc. ICASSP 1985, Tampa, Florida.

Kates, J.M. (1991). A time-domain digital cochlear model. IEEE Trans Sig Proc, 39(12), 2573 - 2592.

Kates, J.M. (1992). On using coherence to measure distortion in hearing aids. J Acoust Soc Am, 91(4), 2236 - 2244.

Leijon, A. (1989). Optimization of hearing-aid gain and frequency response for cochlear hearing losses (Ph.D. thesis). Technical report no. 189. Chalmers University of Technology, Göteborg, Sweden.

Leijon, A. (1990). Hearing Aid Gain for Loudness-Density Normalization in Cochlear Hearing Losses with Impaired Frequency Resolution. Ear and Hearing, 12(4), 242 - 250.

Levitt, H., Cudahy, E., Hwang, W., Kennedy, E., and Link, C. (1987a). Towards a general measure of distortion. J Rehab Res and Dev., 24(4), 7 - 19.

Levitt, H., Sullivan, J.A., Neuman, A.C., Rubin-Spitz, J.A. (1987b). Experiments with a programmable master hearing aid. J Rehab Res and Dev., 24(4), 29 - 54.

Lyon, R.F. (1982). A computational model of filtering, detection and compression in the cochlea. Proc. ICASSP 1982, 1282 - 1285.

Nielsen, Lars Bramsløw (1992). Subjective evaluation of sound quality for normal-hearing and hearing-impaired listeners. Internal report no. 43-8-1, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 51, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

Nielsen, Lars Bramsløw (1993a). An Auditory Model with Hearing Loss. Internal report no. 43-8-2, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 52, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

Nielsen, Lars Bramsløw (1993b). A Neural Network Model for Prediction of Sound Quality. Internal report no. 43-8-3, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 53, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

Nielsen, Lars Bramsløw (1993c). Objective Scaling of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners (this report). Internal report no. 43-8-4, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 54, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

Paillard, B., Mabilleau, P., Morisette, S. and Soumagne, J. (1992). PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals. J Audio Eng. Soc., 40 (1/2), 21 - 31.

Punch, J.L., Montgomery, A.A., Schwartz, D.M., Walden, B.E., Prosek, R.A. and Howard, M.T. (1980). Multi-dimensional scaling of quality judgments of speech signals processed by hearing aids. J Acoust Soc Am, 68(2), 458 - 466.

Quackenbush, S.R., Barnwell, T.P. & Clements, M.A. (1988). Objective measures of speech quality. Prentice-Hall, New Jersey.

Report 1. See Nielsen (1992).

Report 2. See Nielsen (1993a)

Report 3. See Nielsen (1993b)

Rydén, T. (1993). A Broadcaster's First Experiences from Applying a Perceptual Model for the Prediction of Perceived Sound Quality. Proc. AES 12th conference, "Perception of Reproduced Sound", Copenhagen, Denmark.

Scharf, B. (1978). Loudness. Chapter 6 in: Handbook of Perception. Vol. IV: Hearing. (eds.: Carterette & Friedman). Academic Press, New York.

Schwartz, D.M., Lyregaard, P.E. and Lundh, P. (1988). Hearing Aid Selection for Severe-to-Profound Hearing Loss. Hearing Journal, 39(2), 13 - 17.

Seneff, S. (1985). Pitch and spectral analysis of speech based on an auditory synchrony model. M.I.T. Technical Report 504, 242 pp.

von Bismarck, G. (1974a). Timbre of steady sounds: a factorial investigation of verbal attributes. Acustica 30, 146 - 159.

von Bismarck, G. (1974b). Sharpness as an attribute of the timbre of steady sounds. Acustica 30, 159 - 172.

Zwicker, E. & Fastl, H (1990). Psychoacoustics - facts and models. Springer, Berlin.

Zwicker, E. & Feldtkeller, R. (1967). Das Ohr als Nachrichtenempfanger. Hirzel, Stuttgart.

13 Appendices.

13.1 Abstract - Report 1.

Nielsen, Lars Bramsløw (1992). **Subjective Evaluation of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners.** Internal report no. 43-8-1, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 51, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

11 hearing-impaired (HI) and 12 normal-hearing (NH) subjects have performed sound quality ratings on 6 perceptual scales (Loudness, Clarity, Sharpness, Fullness, Spaciousness and Overall judgment). The signals for the rating experiment consisted of running speech and music with or without background noise. These signals were processed in various configurations of filtering, clipping, and compression to form a total of 64 stimuli. Each stimulus was presented monaurally over headphones and rated three times during successive visits. The stimuli for HI listeners were amplified using the POGO fitting rule.

One major purpose of the experiment was to provide data for an objective measure of sound quality. The obtained data covered a large range on the rating scales, which represented different underlying perceptual scales.

All subjects performed the rating task in a satisfactory manner, but the normal-hearing group was slightly more reliable. There were significant differences between stimuli and between subjects, with stimuli affecting the ratings the most. Normal-hearing and hearing-impaired subjects showed similar trends, but normal-hearing listeners were generally more sensitive, i.e. covered a larger range on each rating scale. Of the chosen signal processing parameters, spectral modifications affected the perceived sound quality the most, but clipping and compression also produced detectable differences.

The perceived sound quality could be described by four underlying perceptual dimensions or, with simpler interpretation, by four of the original rating scales.

The two subject groups agreed in their interpretation of the rating scales, and were almost identical in their use of the scales. Based on this, the rating scales were not considered absolute scales.

13.2 Table of contents - Report 1.

1. Introduction. 1
1.1 Definition of sound quality. 1
1.2 Sound quality assessment. 5
1.3 Literature review. 8
2. Purpose 17
3. Method. 21
3.1 Signals. 21
3.2 Stimulus generation 22
3.2.1 Signal processing. 22
3.2.2 Hearing loss compensation. 27
3.3 Experimental setup. 27
3.4 Subjects. 28
3.5 Rating scales and procedure. 32
3.6 Experimental design. 36
3.7 Experiment protocol. 37
3.8 Data treatment. 39
4. Results. 41
4.1 Individual results. 41
4.2 Group results. 46
4.3 Signal processing effects. 55
4.4 Rating scales and perceptual dimensions. 69
4.5 Demonstration tape. 76
5. Discussion. 79
6. Conclusion. 85
7. Literature list. 87
Appendices. 91
A1: Signals and processing 91

A2: Experimental design. 96
A3: Signal processing software: Documentation. 105
A4: Experimental equipment and set-up. 109
A5: Subject summary. 112
A6: Subject instruction and sample data. 113
A7: Statistical models and results. 118

13.3 Abstract - Report 2.

Nielsen, Lars Bramsløw (1993a). **An Auditory Model with Hearing Loss.** Internal report no. 43-8-2, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 52, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

An auditory model based on the psychophysics of hearing has been developed and tested. The model simulates the normal ear or an impaired ear with a given hearing loss. Based on reviews of the current literature, the frequency selectivity and loudness growth as functions of threshold and stimulus level have been found and implemented in the model.

The auditory model was verified against selected results from the literature, and it was confirmed that the normal spread of masking and loudness growth could be simulated in the model. The effects of hearing loss on these parameters was also in qualitative agreement with recent findings. The temporal properties of the ear have currently not been included in the model.

As an example of a real-world application of the model, loudness spectrograms for a speech utterance were presented. By introducing hearing loss, the speech sounds became less audible and less detailed, a problem that linear amplification did not solve properly. This demonstrated how the model could be used for hearing aid development and evaluation.

13.4 Table of contents - Report 2.

1. Introduction. 9
2. Literature review. 11
2.1 Cochlear models. 11
2.2 Physiological measurements. 13
2.3 Cochlear modeling problems. 16
2.4 Auditory models. 17
2.5 Psychophysical measurements. 18

3. Model description. 20
3.1 Model structure. 20
3.2 Power spectrum calculation. 22
3.3 Equalizations and coupler corrections. 23
3.4 Auditory filter bank. 28
3.4.1. Filter shape as a function of level. 33
3.4.2. Filter shape as a function of hearing loss. 36
3.4.3. Filter shape as a function of level and hearing loss. 41
3.5 Loudness function. 46
3.5.1. As a function of level and threshold. 46
3.5.2. Loudness summation in hearing-impaired listeners. 51
3.6 Temporal processing. 52
4. Verification. 53
4.1 Test design and stimuli. 53
4.2 Frequency selectivity. 53
4.2.1. Excitation patterns, pure tones. 53
4.2.2. Noise signals. 56
4.2.3. Impaired frequency selectivity. 58
4.3 Loudness. 61
4.3.1. Loudness growth in normal and impaired hearing. 61
4.3.2. Equal loudness level contours. 67
4.4 Temporal resolution. 68
5. Processing of real-world signals. 69
5.1 Perception of speech sounds. 69
5.2 Performance and future improvements. 71
6. Conclusion. 73
7. References. 75
8. Appendices. 81
8.1 User manual. 81
8.1.1. Input parameter file format. 82
8.1.2. Command-line usage. 87
8.2 Proposed UCL-encoding. 90

13.5 Abstract - Report 3.

Nielsen, Lars Bramsløw (1993b). **A Neural Network Model for Prediction of Sound Quality.** Internal report no. 43-8-3, Oticon Research Unit, Snekkersten, Denmark. Also published as: Report no. 53, The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark.

An artificial neural network structure has been specified, implemented and optimized for the purpose of predicting the perceived sound quality for normal-hearing and hearing-impaired subjects. The network was implemented by means of commercially available software and optimized to predict results obtained in subjective sound quality rating experiments based on input data from an auditory model.

Various types of input data and data representations from the auditory model were used as input data for the chosen network structure, which was a three-layer perceptron. This network was trained by means of a standard backpropagation procedure and tested on selected stimuli from the subjective rating experiment. The best results were obtained with an additional input to the network, identifying the listener, and thus allowing different states for each subject.

The performance with previously unseen test was evaluated for two types of test set extracted from the complete data set. With a test set consisting of mixed stimuli, the prediction error was only slightly larger than the statistical error in the training data itself. Using a particular group of stimuli for the test set, there was a systematic prediction error on the test set. The overall concept proved functional, but further testing with data obtained from a new rating experiment is necessary to better assess the utility of this measure.

The weights in the trained neural networks were analyzed to qualitatively interpret the relation between the physical signal parameters and the subjectively perceived sound quality. No simple objective-subjective relationship was evident from this analysis.

13.6 Table of contents - Report 3.

1. An introduction to neural nets. 9
2. Speech-related applications. 15
3. Scope and purpose of neural net application. 21
4. Model architecture. 23
4.1 Input data representation. 24
4.1.1. Spectral data reduction. 28
4.1.2. Temporal data reduction. 30
4.1.3. Other inputs. 31
4.2 Output data representation. 33
4.3 Neural network implementation. 34
5. Model training and testing principles. 37
5.1 Training algorithm. 37
5.2 Training performance. 38
5.3 Test facts. 39
5.4 Test performance. 40
6. Training sessions. 43
6.1 Training and test schedule. 43
6.2 Single subject. 45
6.3 Subject group (Normal hearing). 50
6.4 Subject group, with subject input. 53
6.4.1. Normal hearing. 54
6.4.2. Hearing impaired. 60
6.5 Test with a class of stimuli. 62
6.5.1. Normal hearing. 63
6.5.2. Hearing impaired. 65
7. Analysis of network weights. 67
8. Discussion. 73
9. Conclusion. 77
10. References. 79
11. Appendices. 83
11.1 Calibration and stimulus levels. 83

11.2 Auditory model parameter files. 84
11.3 Example of auditory model output correlation. 86
11.4 List of stimuli and test sets. 88